

UNIVERSIDAD CARLOS III DE MADRID

DEPARTAMENTO DE BIBLIOTECONOMÍA Y DOCUMENTACIÓN



TRABAJO FIN DE MÁSTER

Aplicación de los principios Linked Open Data a la lista de encabezamientos de materia de la Biblioteca de la Universidad Politécnica de Madrid

Alumno: Rafael Ávila Alonso

Tutora: Dra. Virginia Ortiz Repiso

Presentación: Septiembre 2014

AGRADECIMIENTOS

Quiero mencionar y reconocer a aquellas personas que han colaborado de modo desinteresado y profesional en la confección de este proyecto.

En primer lugar agradecer al Servicio de Biblioteca Universitaria de la Universidad Politécnica de Madrid, por su buena disposición y asistencia, fundamentalmente a través de María José Carrillo, e Isabel Domecq. Mencionar también a Alejandro Martínez, compañero de la Biblioteca del Campus Sur, por su amable atención y por esos extensos diálogos bibliotecarios que han sido muy orientativos para mí.

Doy las gracias también a mi tutora, Virginia Ortiz, por dirigir con paciencia y amable orientación este proyecto, y por todos los años en los que me ha impartido docencia de modo riguroso pero a la vez cercano.

Final y especialmente, agradezco a mi compañera de viaje Pilar Jesús Muñoz, su ayuda sin límites. Valoro especialmente sus palabras sinceras y certeras, y el no menos apreciado apoyo en las correcciones y el estilo.



Esta obra está licenciada bajo una [Licencia Creative Commons Atribución 4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

La ciencia más útil es aquella cuyo fruto es el más comunicable.

Leonardo da Vinci

RESUMEN

Este proyecto pretende encuadrarse en los procedimientos establecidos para la gestión y publicación de vocabularios de valor de bibliotecas en la Web de datos mediante tecnologías Linked Data. El objetivo fundamental es establecer procesos útiles para la migración de lenguajes documentales tradicionales a vocabularios en formato digital, definidos para una reutilización libre del contenido y de modo combinado con otros vocabularios en un contexto de interoperabilidad y fomento del multilingüismo. Se pretende con ello aprovechar la riqueza descriptiva que estos vocabularios ofrecen en conjunto, muy necesaria para la descripción y organización del conocimiento digital. Para conseguirlo se efectúa una migración desde listas de encabezamientos de materia a tesauros en su expresión más novedosa, como instrumentos de organización conceptual y plataformas para la interoperabilidad de vocabularios, siguiendo las pautas normativas de la ISO 25964 1 y 2 e intentando no perder la naturaleza propia de los vocabularios de materia de la Biblioteca de la Universidad Politécnica de Madrid.

Linked Data ha sido el instrumento que ha permitido obtener los resultados esperados. Sus tecnologías han dado soporte a las necesidades de expresión semántica del vocabulario convirtiéndolo en un producto interoperable apto para su procesado automático por aplicaciones semánticas. La adición de licencias abiertas lo coloca además en la categoría de datos vinculados y abiertos, lo que supone abrir las puertas a su libre reutilización

El método de acercamiento al tema del proyecto ha consistido en contextualizar ámbitos de información cada vez más específicos, ayudando a la comprensión y consecución del objetivo principal de este trabajo en un acercamiento progresivo a las técnicas que han facilitado su desarrollo y han concluido en su publicación.

El seguimiento de todo el proceso ha demostrado claramente dos consecuencias básicas: que es posible alinear los productos bibliotecarios con las tecnologías de descripción semántica, incluso contando con escasos recursos para ello, y que la acometida de proyectos de esta índole necesita de la formación de grupos de trabajo multidisciplinares si se quieren conseguir unos mínimos niveles de calidad.

PALABRAS CLAVE: Linked Open Data, Web semántica, Vocabularios, Tesauros, Encabezamientos de materia, Web de datos, Bibliotecas, Organización del conocimiento, SKOS.

DESCRIPCIÓN SEMÁNTICA

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix dcmitype: <http://purl.org/dc/dcmitype/>
@prefix skos: <http://www.w3.org/2004/02/skos/core#>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
```

```
<http://www.a falta de URL> a dcterms:BibliographicResource ;
    dc:title "Aplicación de los principios Linked Open Data a la
    lista de encabezamientos de materia de la Biblioteca de la
    Universidad Politécnica de Madrid" ;
    dcterms:creator _:autor ;
    dcterms:contributor _:tutor ;
    dcterms:publisher "Universidad Carlos III de Madrid" ;
    dcterms:type "Text" # http://dublincore.org/documents/dcmi-type-
vocabulary/#
    dcterms:created ;
    dcterms:license <http://creativecommons.org/licenses/by/4.0/>;
    dcterms:provenance _:deprocedencia ;
    dcterms:bibliographicCitation ;
    dcterms:subject
    http://id.loc.gov/authorities/subjects/sh2013002090 ,
    http://id.loc.gov/authorities/subjects/sh2002006395 ,
    http://id.loc.gov/authorities/subjects/sh2002000569 ,
    http://id.loc.gov/authorities/subjects/sh97007353 ,
    http://id.loc.gov/authorities/subjects/sh2003010124 ,
    http://id.loc.gov/authorities/subjects/sh85129426 ,
    http://id.loc.gov/authorities/subjects/sh85134827 ;
    dcterms:abstract "Proyecto para la publicación de los
vocabularios de materias de la Biblioteca de la Universidad Politécnica
de Madrid bajo tecnologías Linked Open Data. El proyecto contempla dos
vías de actuación fundamentales: la migración de un vocabulario de
materias tradicional en un tesoro conceptual y el modelado y
publicación de dicho tesoro en Linked Open Data." .

_:autor foaf:name "Rafael Ávila Alonso" ;
    foaf:mbox "rafael.avila@xupm.es" .
_:contributor dcterms:identifier <http://viaf.org/viaf/65755283>
    foaf:name "Virginia Ortiz Repiso" ;
    foaf:homepage
<http://portal.uc3m.es/portal/page/portal/biblioteconomia_documentacion
/profesores/virginia> .
_:deprocedencia dcterms:rightsHolder "Rafael Ávila" ;
    dcterms:dateSubmitted "21/09/2014" ;
    dcterms:rightsHolder "Universidad Carlos III de Madrid" ;
    dcterms:dateAccepted "xx/xx/xx" .
```

SUMARIO

| | | |
|-------|--|----|
| 1 | OBJETO, METODOLOGÍA Y FUENTES..... | 1 |
| 1.1 | INTRODUCCIÓN | 1 |
| 1.2 | OBJETO Y JUSTIFICACIÓN | 2 |
| 1.3 | METODOLOGÍA..... | 3 |
| 1.4 | FUENTES | 4 |
| 2 | LINKED DATA, UNA PERSPECTIVA GLOBAL | 7 |
| 2.1 | PANORAMA GLOBAL SOBRE DATOS..... | 7 |
| 2.2 | BIG DATA | 7 |
| 2.2.1 | DEFINICIÓN..... | 7 |
| 2.2.2 | ESTRUCTURA DE DATOS Y RETOS TÉCNICOS..... | 8 |
| 2.2.3 | EFFECTOS DE BIG DATA: IMPACTO ECONÓMICO Y SOCIAL | 9 |
| 2.2.4 | BIG DATA: APERTURA DE DATOS Y APLICACIÓN DE TECNOLOGÍAS SEMÁNTICAS | 11 |
| 2.3 | OPEN DATA..... | 13 |
| 2.3.1 | ¿QUÉ SIGNIFICA OPEN DATA?..... | 13 |
| 2.3.2 | ¿QUÉ CONSECUENCIAS TIENE LA APERTURA DE DATOS?..... | 15 |
| 2.3.3 | OPEN GOVERNMENT: EL MOTOR DEL CAMBIO | 17 |
| 2.3.4 | MARCO LEGAL DEL OPEN GOVERNMENT DATA | 19 |
| 2.3.5 | ¿CÓMO ABRIR LOS DATOS?..... | 21 |
| 2.3.6 | LOS PROBLEMAS DE LA APERTURA DE DATOS..... | 23 |
| 2.4 | DATOS ABIERTOS DE INVESTIGACIÓN | 25 |
| 2.4.1 | UNA APROXIMACIÓN AL TEMA..... | 25 |
| 2.4.2 | DATA SHARING | 25 |
| 2.4.3 | EL PROGRAMA HORIZON 2020..... | 30 |
| 2.4.4 | CURACIÓN DE DATOS DE INVESTIGACIÓN | 32 |
| 2.4.5 | LOS ACTORES DE LA GESTIÓN DE DATOS. PROFESIONALES DE LA INFORMACIÓN E INVESTIGADORES | 34 |
| 2.5 | ¿QUÉ ES LINKED DATA?..... | 36 |
| 2.5.1 | UNA PERSPECTIVA GENERAL..... | 37 |
| 2.5.2 | TECNOLOGÍAS LINKED DATA | 38 |
| 2.5.3 | MODELO DE SERVICIO DE DATOS EN LINKED DATA..... | 61 |
| 2.5.4 | CICLO DE VIDA DE LOS DATOS BAJO TECNOLOGÍAS LINKED DATA..... | 62 |

| | | |
|-------|---|-----|
| 3 | LINKED OPEN DATA – GLAM (GALLERIES, LIBRARIES, ARCHIVES, MUSEUMS) | 66 |
| 3.1 | LINKED DATA Y LAS INSTITUCIONES DEL PATRIMONIO CULTURAL | 66 |
| 3.1.1 | LIBRARY LINKED DATA INCUBATOR GROUP FINAL REPORT..... | 66 |
| 3.1.2 | SITUACIÓN ACTUAL DE LAS BIBLIOTECAS | 68 |
| 3.1.3 | BIBLIOTECAS Y AGENTES EXTERNOS DE LA CULTURA..... | 70 |
| 3.1.4 | CATEGORÍAS DE DATOS BIBLIOTECARIOS | 71 |
| 3.2 | ESTÁNDARES BIBLIOGRÁFICOS DISPONIBLES EN LA WEB SEMÁNTICA | 72 |
| 3.2.1 | DUBLIN CORE..... | 74 |
| 3.2.2 | FRBR Y DATOS VINCULADOS | 75 |
| 3.2.3 | PERSPECTIVAS DEL FORMATO MARC..... | 78 |
| 3.2.4 | ADAPTACIÓN DE ISBD A LINKED OPEN DATA..... | 79 |
| 3.2.5 | RDA Y LINKED DATA..... | 81 |
| 3.2.6 | SCHEMA ORG..... | 84 |
| 3.2.7 | LINKED OPEN DATA ENABLED BIBLIOGRAPHIC DATA 2.0 LODE-BD..... | 86 |
| 3.2.8 | BIBFRAME..... | 88 |
| 3.3 | CASOS DE USO | 94 |
| 3.3.1 | LIBRARY LINKED DATA INCUBATOR GROUP REPORT: USES CASES..... | 94 |
| 3.3.2 | EUROPEANA | 96 |
| 3.3.3 | BIBLIOTECA NACIONAL DE ESPAÑA. DATOS BNE 2.0 | 101 |
| 3.4 | BÚSQUEDA SEMÁNTICA DE LA INFORMACIÓN. HERRAMIENTAS DE DESCUBRIMIENTO..... | 103 |
| 3.5 | UN ANÁLISIS CRÍTICO DE LAS TECNOLOGÍAS LINKED DATA EN BIBLIOTECAS..... | 108 |
| 4 | REPRESENTACIÓN DE SISTEMAS DE ORGANIZACIÓN DEL CONOCIMIENTO | 111 |
| 4.1 | SISTEMAS PARA LA ORGANIZACIÓN DEL CONOCIMIENTO | 111 |
| 4.1.1 | TESAUROS Y ENCABEZAMIENTOS DE MATERIA. KOS PARA LA WEB DE DATOS | 112 |
| 4.1.2 | VOCABULARIOS DE VALORES | 114 |
| 4.2 | SIMPLE KNOWLEDGE ORGANIZATIONN SYSTEM: SKOS..... | 116 |
| 4.2.1 | ESTRUCTURA GENERAL DEL MODELO SKOS | 116 |
| 4.2.2 | ESQUEMA DEL MODELO DE DATOS DE SKOS..... | 117 |
| 4.2.3 | SKOS EXTENSION FOR LABELS. SKOS-XL | 122 |
| 4.2.4 | EVOLUCIÓN DEL MODELO SKOS..... | 124 |
| 4.2.5 | MAPEOS CON SKOS | 125 |
| 4.3 | AJUSTE DE LA NORMA ISO 25964 CON SKOS | 127 |
| 4.4 | MADS..... | 129 |

| | | |
|-------|--|-----|
| 4.4.1 | ELEMENTOS BÁSICOS DE MADS | 129 |
| 4.4.2 | GESTIÓN DE LA PRECOORDINACIÓN CON MADS | 130 |
| 4.5 | VOCABULARIOS DE MATERIAS. ANÁLISIS DE CASOS DE USO..... | 132 |
| 4.5.1 | LIBRARY OF CONGRESS SUBJECT HEADINGS | 132 |
| 4.5.2 | RAMEAU | 134 |
| 4.5.3 | SCHLAGWORT NORMDATEI (SWD) | 134 |
| 4.5.4 | NUEVO SOGGIETTARIO..... | 135 |
| 4.6 | VINCULACIÓN ENTRE VOCABULARIOS | 136 |
| 5 | VOCABULARIO DE MATERIAS DE LA BIBLIOTECA DE LA UNIVERSIDAD POLITÉCNICA DE MADRID (BUPM) BAJO LAS DIRECTRICES Y RECOMENDACIONES DE LINKED OPEN DATA..... | 138 |
| 5.1 | PLANIFICACIÓN GENERAL..... | 138 |
| 5.1.1 | ANÁLISIS DE LA COMUNIDAD DE USUARIOS..... | 138 |
| 5.1.2 | ANÁLISIS DE RECURSOS..... | 139 |
| 5.2 | IMPLEMENTACIÓN DEL VOCABULARIO..... | 143 |
| 5.2.1 | OBTENCIÓN DE LOS DATOS PARA LA CONSTRUCCIÓN DEL VOCABULARIO | 144 |
| 5.2.2 | ANÁLISIS DE LOS REGISTROS DE MATERIA..... | 144 |
| 5.2.3 | DEFINICIÓN DEL DOMINIO DEL VOCABULARIO Y DE SU ESTRUCTURA GENERAL..... | 146 |
| 5.2.4 | ESTABLECIMIENTO DE UN MODELO DE DATOS | 147 |
| 5.2.5 | SELECCIÓN DE LOS CONCEPTOS DEL PROTOTIPO | 151 |
| 5.2.6 | CONTROL DEL VOCABULARIO | 152 |
| 5.2.7 | GESTIÓN DE LAS RELACIONES | 162 |
| 5.3 | PREPARACIÓN DEL VOCABULARIO PARA LA INTEROPERABILIDAD | 168 |
| 5.3.1 | IDENTIFICACIÓN DE DATOS. INTERNATIONALIZED RESOURCE IDENTIFIERS | 168 |
| 5.3.2 | ESTABLECIMIENTO DE RELACIONES SEMÁNTICAS CON OTROS TESAUROS | 170 |
| 5.4 | SKOSIFICACION | 185 |
| 6 | PUBLICACIÓN DE LA LISTA DE ENCABEZAMIENTOS DE MATERIA DE LA BUPM EN LINKED OPEN DATA 193 | |
| 6.1 | PUBLICACIÓN DE DATASETS..... | 193 |
| 6.1.1 | PLATAFORMAS DE PUBLICACIÓN. PUBLICACIÓN EN POOLPARTY | 194 |
| 6.1.2 | PUBLICACIÓN EN THE DATAHUB | 198 |
| 6.1.3 | ESTRUCTURA DE FICHEROS Y RECURSOS QUE INTEGRAN LA PUBLICACIÓN DEL PROYECTO 199 | |
| 6.2 | VOCABULARIOS DE METADATOS DESCRIPTIVOS PARA LINKED DATASETS | 200 |
| 6.2.1 | VOCABULARY OF INTERLINKED DATASETS (Void)..... | 200 |

| | | |
|-------|--|-----|
| 6.2.2 | DATA CATALOG VOCABULARY (DCAT) | 202 |
| 6.2.3 | PUBLICACIÓN DE DATOS DE LAS ADMINISTRACIONES PÚBLICAS. AJUSTE CON LA NORMA TÉCNICA DE INTEROPERABILIDAD | 205 |
| 6.2.4 | VOCABULARIO PARA LA DESCRIPCIÓN DE LA PROCEDENCIA. PROVENANCE..... | 206 |
| 7 | CONSUMO DE DATOS Y PRESERVACIÓN LINKED DATA | 210 |
| 7.1 | SPARQL PROTOCOL AND RDF QUERY LANGUAGE | 211 |
| 7.2 | LICENCIAS PARA DATOS VINCULADOS | 214 |
| 7.2.1 | Aspectos generales para la asignación de licencias | 214 |
| 7.2.2 | Marco jurídico general | 216 |
| 7.2.3 | Asignación de la licencia al proyecto Tesoros Materias BUPM..... | 217 |
| 7.3 | PRESERVACIÓN EN LINKED DATA..... | 218 |
| 7.3.1 | ASPECTOS GENERALES DE LA PRESERVACIÓN DE DATOS | 218 |
| 7.3.2 | PRESEVACIÓN DE DATOS VINCULADOS | 219 |
| 7.3.3 | PRELIDA | 220 |
| 8 | CONCLUSIONES | 227 |
| 9 | REFERENCIAS..... | 231 |

ÍNDICE DE FIGURAS

| | |
|---|-----|
| Figura 1: Crecimiento de la economía europea por el uso de Big Data y Open Data..... | 11 |
| Figura 2: Índice ePSI referenciado al empleo de Open Data y el nivel de reutilización de la información del sector público..... | 16 |
| Figura 3: Universo de los datos de investigación..... | 22 |
| Figura 4: Esquema del proceso de curación de datos de investigación | 29 |
| Figura 5: Representación gráfica de una tripleta RDF | 42 |
| Figura 6: Nuevas posibilidades de serialización RDF..... | 45 |
| Figura 7: Representación mediante grafos de una declaración RDF con nodo en blanco | 49 |
| Figura 8: Estructura de OWL 2 | 57 |
| Figura 9: Descripción mediante triples de la obra Weaving the Web de Tim Berners Lee | 71 |
| Figura 10: Esquema básico del modelo de entidades y relaciones FRBR | 75 |
| Figura 11: Modelo de RDF de alto nivel para registros bibliográficos | 76 |
| Figura 12: Modelo de datos LODE-BD 2.0..... | 85 |
| Figura 13: Modelo de datos de BIBFRAME | 91 |
| Figura 14: Esquema de clases del modelo de datos de Europeana..... | 96 |
| Figura 15: Esquema de propiedades del modelo de datos de Europeana | 97 |
| Figura 16: Ontología de la Biblioteca Nacional de España..... | 102 |
| Figura 17: Modelo de datos del catálogo LIBRIS..... | 106 |
| Figura 18: Elementos básicos del modelo de datos SKOS | 116 |
| Figura 19: Utilización Library of Congress Classification Number..... | 132 |
| Figura 20: Relación de mapeo entre propiedades de diferentes vocabularios | 135 |
| Figura 21: Secuencia de preferencia de fuentes para nuevos conceptos | 153 |
| Figura 22: Modelo de datos del vocabulario de materias BUPM | 154 |
| Figura 23: Estructura de mapeo en hub..... | 178 |
| Figura 24: Proceso de datos del gestor de tesauros Poolparty | 199 |

| | |
|---|-----|
| Figura 25: Detalle de registro del tesoro de materias BUPM | 200 |
| Figura 26: Detalle de visualización de conceptos en Tesoro BUPM | 201 |
| Figura 27: Modelo de datos Data Catalog Vocabulary..... | 207 |
| Figura 28: Proceso de interacción de los agentes participantes en la cadena de datos de las administraciones públicas | 209 |
| Figura 29: Modelo de datos Provenance | 210 |
| Figura 30: Detalle de ecuación de búsqueda en cliente SPARQL Endpoint..... | 215 |
| Figura 31: Resultados de la búsqueda de la figura 30 | 216 |
| Figura 32: Estrategia de empaquetado para la preservación | 228 |

ÍNDICE DE TABLAS

| | |
|--|-----|
| Tabla 1: Modelo tabular de tripleta RDF | 41 |
| Tabla 2: Principales vocabularios semánticos..... | 43 |
| Tabla 3: Principales tipos de datos para solicitudes de contenido | 60 |
| Tabla 4: Diferentes estándares en la estructura de LOD en bibliotecas..... | 73 |
| Tabla 5: Ajuste de los vocabularios ISO 25964 y SKOS..... | 127 |
| Tabla 6: Principales propiedades para el mapeo de vocabularios..... | 136 |
| Tabla 7: Gestión de actualizaciones en los mapeos..... | 188 |

1 OBJETO, METODOLOGÍA Y FUENTES

1.1 INTRODUCCIÓN

El manejo eficiente de la información ha sido y es un objetivo natural de las bibliotecas. Con la explosión informativa, se ha perdido el control que antaño se tenía sobre los recursos informativos, y los productos que se utilizaban para su organización se muestran poco flexibles para trabajar con eficacia en el nuevo contexto. Ordenar la información y el conocimiento generado sigue siendo necesario en la Web, pero su realización práctica no es fácil, la heterogeneidad y el entorno en constante cambio es el sustrato con el que trabajar y las herramientas tradicionales tienen defectos de ajuste.

Podría decirse que las listas de encabezamientos de materia han sido consideradas la aristocracia de los lenguajes documentales. La calidad de sus características, siempre ha sido reconocida en los ámbitos profesionales, siendo utilizadas con éxito para la indización y la recuperación de la información durante años. Pero como todo producto de estructura compleja, muestra cierta inflexibilidad a la hora de afrontar su integración en los nuevos contextos tecnológicos.

Los tesauros conceptuales, en la nueva definición de la norma ISO 25964, son herramientas ya definidas para el trabajo en entornos automatizados y preparadas para establecer un rico mapa de relaciones entre vocabularios. Transformar listas de encabezamientos de materia en tesauros es una estrategia correcta de adaptación de vocabularios a contextos más flexibles, pero ese cambio no es suficiente.

La dificultad que se plantea es por un lado, saber qué tipo de conversión es la correcta para la migración entre diferentes sistemas para la organización del conocimiento y por otro, cómo hacer este producto más útil a la comunidad.

El primer problema que hay que resolver es cómo conseguir un traslado correcto de las características semánticas del vocabulario de materias a su edición definitiva en forma de tesauro. Para ello se cuenta con herramientas poderosas como las normas para la construcción de tesauros conceptuales y el estudio de casos de uso en otros centros que han efectuado esta misma transición.

El segundo problema es cómo extraer el vocabulario de los silos de datos de la biblioteca. Se trata de exportar el vocabulario a la Web de datos, permitiendo su uso compartido y libre a la comunidad, vinculándolo para su enriquecimiento, con otros vocabularios de materias en diferentes idiomas.

El descenso desde lo más general de la teoría de los datos a las concreciones y resultados prácticos, ha permitido un paulatino acercamiento a una temática de gran complejidad y ha ayudado a comprender las principales consecuencias derivadas de este proyecto. Se requiere un aumento

de la estandarización y la creación de procedimientos comunes en los que se colabore internacionalmente. Se precisa que los agentes que gestionen datos coordinen y sumen sus proyectos, evitando desarrollos unilaterales, pues la mejor interoperabilidad es la que no hace falta. La automatización y usabilidad de las herramientas debe mejorarse, acercando su uso personal no experto. La inclusión del multilingüismo en cualquier proyecto semántico, proceso vital para captar la riqueza cultural de las diferentes áreas idiomáticas. Y, finalmente, la necesidad de integración incondicional de los servicios de información en la “ordenación” del conocimiento digital.

En definitiva, se abre una nueva posibilidad para los servicios de información, cubrir los espacios que nos corresponden es un objetivo prioritario, en otro caso, otros ocuparán ese lugar.

1.2 OBJETO Y JUSTIFICACIÓN

Si se tuviera que definir con una sola palabra el objetivo de este proyecto, sin duda se elegiría la reutilización. Efectivamente, el propósito de este trabajo es reutilizar un vocabulario producido y gestionado con las herramientas tradicionales, proyectando su migración a estructuras conceptuales más flexibles y adecuadas al contexto informativo digital, y permitiendo el libre aprovechamiento del mismo.

Se puede desglosar este objetivo general en otros más específicos. Así, verificar un proceso correcto para la migración de vocabularios, acogiendo a normas y estándares que estructuren y guíen los procesos a seguir. Conseguir un índice de calidad adecuado en la transformación y en el producto final, intentando mantener en lo posible las características específicas de las LEM, pero desarrollando su sistema de relaciones, fundamentalmente, asociativas y jerárquicas. Representar el vocabulario en una interfaz abierta que permita su libre utilización y consulta. Preparar el vocabulario para su expresión semántica a través de Linked Open Data, con lo que se persigue adoptar un formato válido y extendido para su libre reutilización y gestión por aplicaciones autónomas. Establecer vínculos semánticos enriquecedores con otros vocabularios de reconocida calidad (LCSH, RAMEAU, etc.), que no sólo incorporen la equivalencia multilingüe, sino la riqueza autóctona de su terminología. Establecer las descripciones adecuadas de los datasets, con metadatos que permitan conocer su contenido, la estructura de mapeos, los datos de procedencia y de preservación. Finalmente incorporar un sistema de preservación ajustada a los procedimientos PRELIDA e incluyendo sistemas de gestión agrupada de la preservación de recursos mediante empaquetamiento semántico.

Justificar un proyecto de esta naturaleza requiere en primer lugar, reconocer que las listas de encabezamientos de materia son vocabularios de gran calidad y con un amplio espectro temático y terminológico. Estas características las convierten en instrumentos adecuados para organizar el conocimiento que se sustancia en el ámbito digital.

Sus peculiares características, permiten identificar contenidos y recursos de muy diferentes maneras, desde la asignación del simple concepto hasta la combinación reglada de elementos en cuya cúspide se encuentran los encabezamientos precoordinaados. Esta configuración posibilita la consecución de altos niveles de especificidad en la categorización del conocimiento, convirtiendo a este tipo de vocabularios en herramientas muy útiles para su utilización en contextos distribuidos y diversos como son los definidos por la Web de datos.

Es por ello que migrar un vocabulario local de materias a Linked Open Data contribuye a aumentar las posibilidades de que los reutilizadores se beneficien de sus características de calidad, a la vez que se visibiliza el trabajo, a veces oculto, de muchos de los profesionales de los centros de información que han dedicado mucho esfuerzo a su construcción.

Además, la inmersión de las bibliotecas en los trabajos con datos puede dejar de ser una opción, y convertirse en una necesidad derivada de la actual escasez presupuestaria. Acometer este tipo de proyectos puede abrir ámbitos profesionales donde desplegar la nueva biblioteca.

1.3 METODOLOGÍA

La metodología aplicada se basa en un proceso híbrido y combinado, basado tanto en el análisis y estudio de los problemas y objetivos planteados, como en un despliegue práctico que permite generar un producto real aunque en fase de prototipo.

El acercamiento a la disciplina de la gestión de vocabularios en un contexto de datos vinculados ha exigido el estudio de múltiples de procesos y normas de aplicación, referidos al tratamiento de terminologías y al modelado semántico, teniendo en cuenta en todo momento las circunstancias referidas al ciclo completo de gestión de datos vinculados.

Dicho análisis ha comenzado con una aproximación a la situación contextual actual tanto de las tecnologías de datos, como su manifestación en centros de información. Con ello se ha tomado contacto con los principales estándares sobre datos vinculados aplicables a las bibliotecas y especialmente a los vocabularios de propiedades y de valores. Seguir esa pauta, ha permitido conocer las fortalezas y debilidades de las tecnologías semánticas y contrastarlas con las capacidades requeridas para efectuar este proyecto. SKOS, a pesar de su carácter descriptivo generalista, es el lenguaje adecuado para la representación semántica. Para completarlo, se ha propuesto la utilización de otras ontologías de modelado de materias como MADS o la introducción de extensión SKOS-XL en el proceso. Esta aproximación analítica también ha verificado las limitaciones de las aplicaciones para la gestión de vocabularios y su expresión en Linked Data, por lo que ha sido necesario, a falta de la posibilidad de desarrollar y programar herramientas específicas, implementar marcado manual para completar, por ejemplo, el sistema de mapeos.

Uno de los motivos para llevar a cabo un análisis combinado ha sido que a la hora de definir los métodos de migración del vocabulario se debían establecer a la par las posibilidades de *skosificación* del mismo, pues, evidentemente, no todos los vocabularios son susceptibles de su modelado con SKOS. Para la creación de IRIs se ha seguido igualmente métodos estandarizados recogidos en la Norma Técnica de Interoperabilidad, introduciendo procesos estables y comprobados, aunque finalmente, las características del gestor y la disponibilidad de recursos han obligado a asumir la imposibilidad de gestionar las IRIs, asignadas de modo automático por la aplicación para la construcción de tesauros.

Verificada la extracción de los datos de materias y su refinado, las pautas para su gestión han sido extraídas de las especificaciones de la Norma ISO 25964 1-2, que incorporan métodos suficientes para el desarrollo o evolución desde un vocabulario de materias a un tesoro y para el establecimiento de la estructura interoperable. Las especificaciones de modelado se han establecido mediante un diagrama UML que abarca, no sólo la estructura propia del vocabulario, sino también el esquema de mapeos. También se han estudiado, los casos más representativos respecto a la publicación de vocabularios semánticos de materias, lo que ha permitido prever las dificultades que podía presentar la migración proyectada.

El aspecto metodológico de este proyecto está definido con mayor amplitud en el capítulo 5.

1.4 FUENTES

La selección de fuentes utilizadas para el desarrollo de este proyecto se ha efectuado en general sobre estándares o recomendaciones de amplia aceptación y bajo la premisa de la calidad y la actualidad que permitan abordar las consecuencias de la obsolescencia tecnológica.

Para el desarrollo contextual, las fuentes utilizadas mayoritariamente han sido artículos de carácter científico, informes de prestigiosas instituciones y recomendaciones de entidades fundamentales en la estandarización de tecnologías semánticas.

Para la aproximación a las tecnologías Linked Data ha sido fundamental el análisis de las recomendaciones del W3C (World Wide Web Consortium) y del OKFN (Open Knowledge Foundation). Ambas organizaciones proponen el empleo de técnicas y lenguajes estandarizados para el desarrollo de productos semánticos, en un contexto de constante desarrollo y evolución de sus productos.

Concretamente y para el modelado del vocabulario de materias se ha utilizado el estándar SKOS (W3C) y sus extensiones, y la ontología MADS desarrollada y mantenida por la Library of Congress. En la siguiente lista se hace referencia a los vocabularios utilizados y la especificación de la recomendación:

- Vocabulario SKOS:
 - SKOS Simple Knowledge Organization System Reference W3C Recommendation 18 August 2009.
 - SKOS Reference appendix B SKOS eXtension for Labels (SKOS-XL).
 - SKOS Simple Knowledge Organization System Primer W3C Working Group Note 18 August 2009.
- RDF 1.1 Concepts and Abstract Syntax W3C Recommendation 25 February 2014.
- Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation 26 November 2008.
- Metadata Authority Description Schema (MADS) Version 2.0 (2012).

El trabajo con los datos desde el punto de vista de su proceso hasta la publicación de datos vinculados, se ha ajustado en lo posible a la Norma Técnica de Interoperabilidad para la Reutilización de Recursos de Información de 19 de febrero de 2013 y la Directiva 2013/37/UE de modificación de la Directiva 2003/98/UE sobre la reutilización de la información del sector público. Concretamente se han seguido las pautas en cuanto a generación de IRIs, los requisitos mínimos de publicación y de descripción de metadatos de los datasets (DCAT).

La gestión del vocabulario de materias ha sido otra de las partes fundamentales del proyecto. Para los procesos generales se han analizado las recomendaciones del programa de vocabularios del The J. Paul Getty Trust (Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works) y las pautas sobre terminología cultural del Programa Linked Heritage para el enriquecimiento de Europeana. Para la migración del vocabulario a su forma como tesoro y el establecimiento de la interoperabilidad semántica se han seguido las directrices de la norma ISO 25964, cuya primera parte se refiere a los tesauros para la recuperación de la información, y la segunda versa sobre la interoperabilidad con otros vocabularios. En la parte de control del vocabulario, las directrices de la norma se han visto asistidas y contrastadas con las pautas internas definidas por la BUPM para su lista de encabezamientos de materia (Biblioteca Universidad Politécnica de Madrid, 2010) y el Manual de Autoridades de la Biblioteca Nacional de España.

Finalmente, para el estudio semántico de relaciones y las propuestas de incorporación de nuevos conceptos, se han utilizado las siguientes fuentes que se dividen en autorizadas o no autorizadas, dependiendo de su mención en las Pautas para listas de encabezamientos de materias de la BUPM:

Autorizadas:

- Catálogo de autoridades de materia de la BNE.
- Encabezamientos de materia de la Library of Congress Subject Headings.
- Tesauro INSPEC
- Tesauro de la UNESCO
- EUROVOC
- Tesauros del CINDOC
- Tesauro AGROVOC
- Materiales no electrónicos de reconocida calidad en su contenido.

No autorizadas:

- Repertorio de autoridades de material de la Biblioteca Nacional de Francia, RAMEAU.
- Directorio de autoridades de materia de la Biblioteca Nacional de Alemania, GND SWD.
- El Tesauro de materias de la Biblioteca Nacional Central de Florencia, Nuovo Soggetario.
- Vocabularios de la Fundación J. Paul Getty.
- Tesauro de la Biblioteca Nacional de Agricultura.

2 LINKED DATA, UNA PERSPECTIVA GLOBAL

2.1 PANORAMA GLOBAL SOBRE DATOS

La Sociedad actual genera, en su actividad cotidiana, una gran cantidad de datos muy heterogéneos. Estos datos pueden tener diversa utilización y origen: pueden ser datos provenientes de la investigación científica, datos de carácter económico, datos derivados del funcionamiento de las administraciones públicas, datos de actividad de las empresas, datos de las instituciones culturales, etc. Las áreas que se van a analizar en este trabajo son cuatro:

1. Big Data, que hace referencia a los datos producidos por las corporaciones en enormes cantidades y con una gran velocidad de producción, a la que se suman los datos producidos masivamente por redes sociales, sensores y dispositivos móviles de todo tipo, etc.
2. Open Data, se centra en cambio, en la apertura de los datos y sus consecuencias.
3. Los datos de investigación que abarcan la estructura factual de la investigación científica.
4. Y Linked Data, cuya principal característica es la interoperabilidad propiciada por la vinculación de datos y su proceso automático. de nuevas utilidades no previstas.

Estos tipos básicos no se estructuran como compartimentos estancos, más bien son conjuntos interrelacionados y complementarios que conforman una nueva web, la Web de los datos. En los siguientes epígrafes se va a analizar con algo más de profundidad, la tipología básica antes descrita.

2.2 BIG DATA

2.2.1 DEFINICIÓN

Estamos inmersos en una verdadera revolución de datos, la Sociedad de la Información sigue evolucionando cada vez con mayor velocidad y ahora se enfrenta a un nuevo reto: una especie de Big Bang de datos. El concepto de Big Data, tal como apunta la consultora Mckensey, (Manyika et al., 2011), hace referencia a datos que por su cuantía, exceden la capacidad de procesamiento, almacenaje, análisis, gestión y captura en un momento determinado. Podemos completar este concepto con el definido por Gartner (2012), cuando afirma que Big data son activos de información en grandes cantidades, recibidos a gran velocidad y definidos en tipos muy variados, los cuales requieren innovadoras fuentes de procesamiento de la información que mejoren su

entendimiento y ayuden en la toma de decisiones. Los datos utilizados en la actualidad no dejan lugar a dudas, se puede cuantificar sobre unos 2,5 millones de Zettabytes (un billón de Gigabytes), con un crecimiento previsto para 2020 de hasta los 100 millones de ZB (García, 2014).

¿A qué se debe esta proliferación de los datos? A los tradicionales productores de datos como las empresas, las administraciones públicas y las corporaciones financieras, se unen ahora los datos emanados de las redes sociales, los dispositivos móviles, los geodatos, los datos transaccionales, los de sensores (también en su propia explosión en el contexto de Internet of Things) etc.

El fenómeno Big Data afecta de modo fundamental al mundo de las nuevas tecnologías y principalmente al de la Información en todos sus ámbitos. En las empresas favorece la toma de decisiones y las políticas comerciales ajustadas a la competencia; en la investigación científica permite el análisis de la enorme cantidad de datos que provienen por ejemplo de la Genética y en el ámbito político también apoya a la toma de decisiones basada en el análisis de datos. Big Data permite, a través de los nuevos avances en software y hardware, el análisis de cantidades masivas de datos extrayendo su valor implícito o el que se deduce en la puesta en relación con su contexto de utilización (Buchholtz, Bukowski, & Śniegocki, 2014; Manyika et al., 2011).

2.2.2 ESTRUCTURA DE DATOS Y RETOS TÉCNICOS

Habitualmente Big Data considera tres atributos esenciales a la hora de caracterizar los datos: volumen, velocidad y variedad de datos.

1. Volumen de datos: Big Data supone de por sí, grandes cantidades de datos recogidos y procesables. El volumen de datos es un atributo relativo, pues grandes cantidades de datos hoy pueden no ser tantos mañana.
2. Velocidad de creación de los datos: los datos se generan con gran rapidez y lo que es más importante, esa velocidad de generación aumenta exponencialmente. Las herramientas de análisis están verificando un aumento en paralelo de su capacidad de proceso, esto permite obtener resultados cercanos al tiempo real y permite actualizaciones constantes de los resultados de los análisis de datos.
3. Variedad de datos: los datos procesados no provienen únicamente de la actividad interna de las organizaciones, existen ahora una multiplicidad de fuentes externas y formatos, que dejan obsoletos los sistemas de análisis tradicionales. Esto permite, dada la riqueza de los datos a analizar, el descubrimiento de nuevo conocimiento.

La propia evolución del tema está determinando la aparición de otros más novedosos como el proceso de calidad de los datos y su valor económico y social. (Buchholtz et al., 2014; Mitchell & Wilson, 2012):

4. Veracidad de los datos: se requiere calidad y fiabilidad para hacer posible su análisis.
5. Valor de los datos: su análisis datos tiene un valor económico y social, que puede beneficiar a las organizaciones y por extensión a la Sociedad.

Uno de los puntos clave del aspecto técnico/tecnológico del Big Data reside en el análisis de datos. Estas técnicas comprenden el uso de la Estadística, la Informática, la Computación avanzada, la Investigación operativa, la Inteligencia artificial etc. La cristalización teórica de estos aspectos ha generado la llamada Ciencia de los Datos, productora de los necesarios y escasos perfiles de científicos de datos (Fox, 2013).

Big Data se enfrenta a ciertos retos tecnológicos que apoyen su implantación masiva y eficaz en un contexto de dinamismo e innovación tecnológica (De Lama, 2012):

1. Nuevas necesidades de almacenamiento no tradicional, especialmente enfocado a los datos desestructurados (bases de datos NoSQL).
2. Nuevas necesidades de proceso, plataformas de computación distribuidas (MapReduce, Hadoop). Capacidades de análisis dinámicas en tiempo real.
3. Nuevos requerimientos y capacidades distribuidas para el análisis de datos. Evolución en las técnicas de minería de datos, en el aprendizaje automático de computadoras y novedosas técnicas de visualización de datos.
4. Adaptación de las competencias necesarias para la generación de nuevos perfiles de científicos de datos.

2.2.3 EFECTOS DE BIG DATA: IMPACTO ECONÓMICO Y SOCIAL

Tal como refiere la consultora McKinsey (Manyika et al., 2011), el paradigma Big Data supone que la innovación y el crecimiento económico no es posible sin la gestión de datos. El valor potencial de los datos puede impulsar la economía y por ende a la Sociedad hacia nuevos niveles desconocidos de prosperidad, tal y como se produjo con los grandes avances tecnológicos del pasado. Big Data, supondrá una mejora de la productividad y de la competitividad de las organizaciones, lo que redundará en mejoras económicas importantes para los consumidores. El coste de extracción de información es cada vez más asumible gracias a las posibilidades del Cloud Computing y del software abierto, verdaderos aceleradores de la innovación tecnológica y la evolución económica.

McKinsey basa su explicación de Big Data en estos puntos fundamentales:

1. Los datos están instalados en el tejido económico, son un valor fundamental en la producción de las organizaciones.
2. Big Data genera valor:
 - a. Transparencia de las administraciones públicas (por ejemplo la obligatoria compartición de datos de investigación, o los datos presupuestarios públicos).
 - b. Segmentación poblacional más ajustada y mejora de la oferta de servicios y productos.
 - c. Innovación y mejora de la producción en las organizaciones.
 - d. Mejoras en el apoyo a las toma de decisiones de los gestores.
3. Mejora de la competitividad de las empresas provocada por un mejor entendimiento del contexto económico derivado del análisis de los datos.
4. Big Data beneficia más a unos sectores que a otros y refiere previsibles ventajas del sector de la Computación y de la Información.
5. Se requiere más talento en los gestores de los datos. Es preciso definir el perfil del científico de datos y la adecuación de sus competencias hacia cualificaciones en disciplinas como las matemáticas, la estadística o la ingeniería informática (Tascón, 2013).
6. Políticas de datos definidas: la cuestión de la propiedad intelectual de los datos y de la protección de datos, transacciones de datos (compras, compartición de datos).
7. Evolución tecnológica que soporte la progresión en el crecimiento del volumen de datos y su complejidad

El informe demosEuropa (Buchholtz et al., 2014) recoge también la importancia de la mejora de los procesos I+D provocada por el análisis de datos. En la figura 1 se muestra el potencial de crecimiento económico europeo en la próxima década por la utilización de Big Data y Open Data

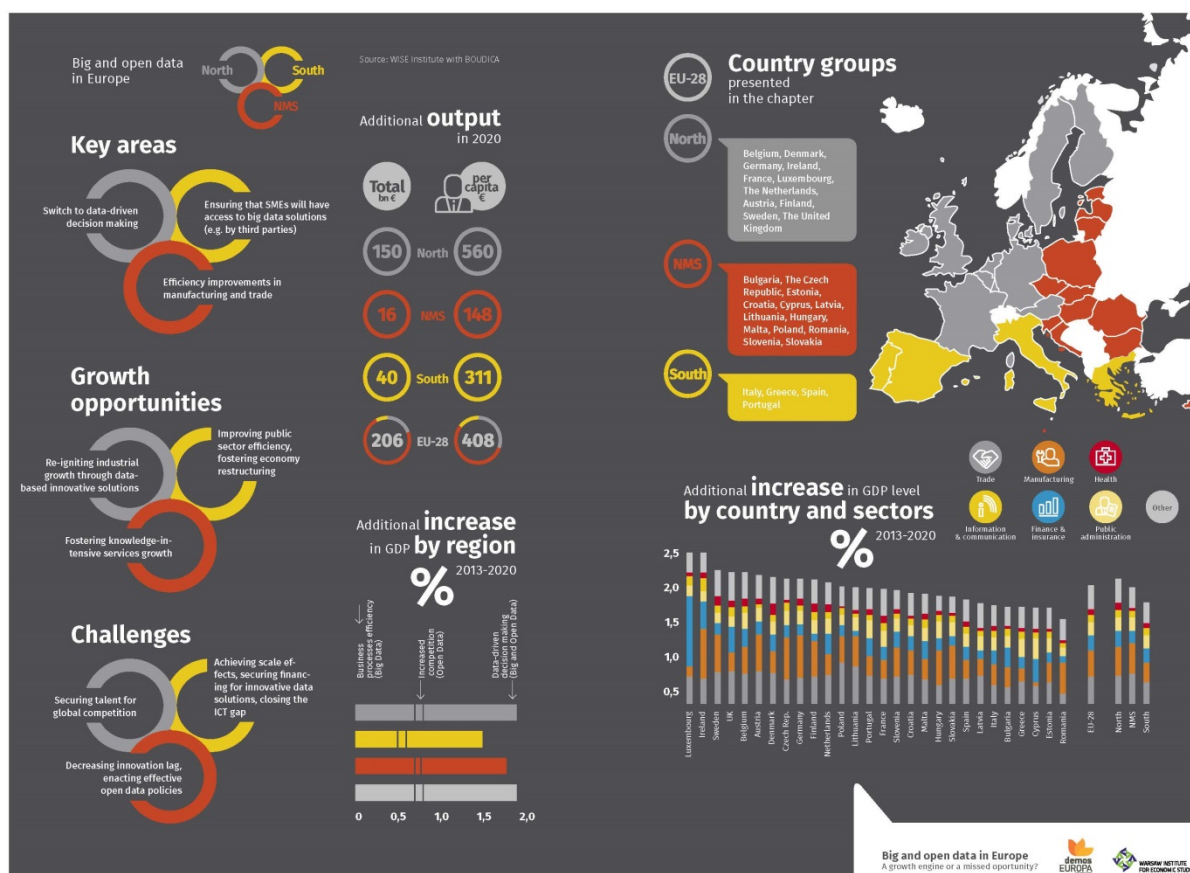


Figura 1 Crecimiento de la economía europea por el uso de Big Data y Open Data. Fuente: Warsaw Institute for Economic Studies. 2014

2.2.4 BIG DATA: APERTURA DE DATOS Y APLICACIÓN DE TECNOLOGÍAS SEMÁNTICAS

Existe un cierto consenso en definir Big Data como conjunto global de compleja integración que abarca en su dimensión a los datos de investigación, los datos abiertos, los datos vinculados, datos de redes sociales, etc. (Shiri, 2014). Todas estas fuentes suponen una mezcolanza heterogénea de datos que responden a filosofías diferentes y con gran diversidad en su estructuración. Su gestión plantea grandes desafíos tecnológicos y metodológicos: el descubrimiento y selección de los datos, su extracción y procesado final, la preservación, la visualización, la posibilidad de acceso a los datos, su estructuración mayor o menor, etc.

La apertura de datos en el contexto Big Data puede beneficiar doblemente a las organizaciones, aumentando la transparencia de su organización y sus productos o servicios, propiciando un

acercamiento a los clientes y presumiblemente generar más beneficios (Ferrer-Sapena & Sánchez-Pérez, 2013). También la apertura de los datos entre consorcios de empresas, puede beneficiar el crecimiento y eliminar las ineficiencias. Del mismo modo las técnicas Big Data pueden servir para la extracción de conocimiento de los grandes conjuntos de datos públicos abiertos, mejorando tanto la eficiencia de las Administraciones, como la percepción de dichas mejoras por parte de los contribuyentes.

Linked Data aporta a Big Data un mayor nivel formal y de estructuración de los datos. Mejora la organización del conocimiento derivada de las técnicas de vinculación de datos y de sus componentes, como los vocabularios controlados en Linked Data, utilizando estándares como SKOS Core que introducen contenido semántico en los análisis de datos y su visualización (Shiri, 2014). Los vocabularios Linked Data son también fundamentales para la unificación de datasets; sin los mapeos correspondientes entre los conjuntos de datos, no parece posible la fusión de los mismos ni su análisis (Fox, 2013).

Según Mitchell & Wilson, en el informe de Fujitsu Services Limited, Linked Data puede aportar ventajas importantes en la macro gestión de datos. Una muy importante es ayudar en entornos corporativos a vincular la nube de datos desestructurados externos, a las BBDD internas corporativas, actuando como filtro integrador y permitiendo conseguir más conocimiento de modo creativo (Hitzler & Janowicz, 2013).

Linked Data converge con los sistemas RESTful: la generación de nodos según la diversa función que los datos desempeñan puede acompañarse de la asignación de un IRI, que puede vincularse con otros nodos y sus respectivos IRIs, formando un grafo que integra muy diversas fuentes de datos, enriquecedoras de los análisis Big Data. Linked Data mejora la validez de los datos, ayudando a que conserven su integridad y a evitar errores, esto es fundamental a la hora de la toma de decisiones en las organizaciones. Los vínculos Linked Data permiten una mayor robustez de la integridad referencial de los datos y apoyan a los respaldos de seguridad de los datos vinculándolos (I. Mitchell & Wilson, 2012).

La aplicación de tecnologías semánticas a Big Data presenta ciertos desafíos importantes (Bizer, Boncz, Brodie, & Erling, 2011):

1. La integración de datos que Big Data supone se produce en un entorno multidisciplinar y por ende de complejo análisis.
2. La Web de datos por un lado y los datos estructurados por otro, como parte del ecosistema Big Data, suponen un desafío de integración de los datos y de su procesamiento.
3. Se requieren más casos de uso de estas tecnologías para experimentar en este contexto con datos vinculados y poder demostrar su valor tanto en la integración de datos como en su vinculación.

En este último punto se podría concluir que Linked Data es un pequeño Big Data de laboratorio donde todas las variables aparecen controladas (Hitzler & Janowicz, 2013). Es aquí donde aparecen algunas dificultades importantes, referidas a la calidad de los datos vinculados, probablemente derivadas de la enorme variedad de fuentes, cada una con sus propios modelos: tripletas defectuosas, sintaxis incorrecta o falta de sistemas de consulta SPARQL endpoints, etc. Evidentemente investigadores y profesionales han de buscar soluciones técnicas y de método que permitan diseñar un sistema de buenas prácticas flexible y ajustable a las diversas circunstancias de publicación de datasets.

2.3 OPEN DATA

2.3.1 ¿QUÉ SIGNIFICA OPEN DATA?

Los datos pueden estudiarse desde muy diferentes puntos de vista, cuando hablamos de Open Data, el foco se establece fundamentalmente en la libertad de utilización de esos datos. En cualquiera de las definiciones que se puedan encontrar en la literatura sobre el tema, las características de los datos abiertos se concentran en la posibilidad de reutilización y redistribución libre, siendo ese grado de libertad definido por una licencia. La libertad de uso debe tender a situaciones con las mínimas restricciones posibles: ni económicas, ni tecnológicas, ni políticas, todo lo contrario, la actividad de apertura de datos se ha de orientar hacia modelos que favorezcan y fomenten potencialmente todas las mejoras que la filosofía Open Data supone: desarrollo y crecimiento económico, aumento de la innovación, mejora del conocimiento interno sobre las organizaciones que abren sus datos, transparencia de las administraciones públicas y establecimiento de un marco social evolucionado hacia nuevas metas en la colaboración y participación de los ciudadanos.

Al igual que con la definición de Big Data, los principios definidores de los datos abiertos son muy variados y se van incrementando con el tiempo y la experiencia. Partiendo de los ocho principios originales generados en la reunión de Sebastopol (California-USA) en 2007 por 30 países defensores del Gobierno abierto, podemos añadir algunos más que ayuden a delimitar qué es un dato abierto (Tauberer, 2007):

1. Todos los datos públicos se pueden poner a disposición de los usuarios excepto los sujetos a limitaciones de privacidad, seguridad y propiedad.
2. Los datos publicados deben estar en su formato más primario, recogidos con el nivel más fino de granularidad y sin transformaciones (formato *raw*).
3. Los datos deben ser publicados lo antes posible, el paso del tiempo les resta valor.

4. Los datos deben tener el grado máximo de accesibilidad, abarcando el más amplio espectro de utilizadores y propósitos de uso.
5. Los datos deben de estar estructurados y puestos a disposición en formatos legibles por máquina.
6. Los datos no deben contener ninguna condición de uso discriminatoria, estando disponibles para cualquiera y sin necesidad de registro.
7. Los datos deben publicarse en formatos no propietarios.
8. Los datos deben ser publicados bajo licencias de uso libre, no debiendo tener limitación alguna en cuanto a propiedad intelectual, de patente, de marca, etc.

La evolución de estos principios y la experiencia de uso han proporcionado los fundamentos para la aparición de dos elementos más, promovidos por la SunLight Foundation (Wonderlich, 2010):

9. Los datos deben ser publicados y actualizados de modo permanente. Los datos publicados deben mantener su calidad y usabilidad a lo largo del tiempo.
10. La tendencia al coste cero es una exigencia de la apertura de datos. Por defecto, el gasto de acceso a los datos debe ser nulo.

Finalmente y como aportaciones más novedosas, la consultora demosEuropa, en su informe sobre el desarrollo e impacto económico de los datos abiertos en Europa, incorpora algunas políticas de índole general que mejoran las condiciones de apertura de los datos (Buchholtz et al., 2014):

11. Los conjuntos de datos deben de abrirse por defecto, sin esperar la petición de los intermediarios u otros agentes que necesiten utilizarlos.
12. El conocimiento de los conjuntos de datos (*datasets*) debe ser descrito mediante la utilización de metadatos. Éstos permitirán a los utilizadores obtener información relevante de dichos conjuntos y permitirán hacer valoraciones de los datos de modo más eficiente.
13. La forma de acceso a los datos debe estar diseñada para un acceso mejorado, favoreciendo retroalimentación por parte de los usuarios de datos y aprovechando esa información para mejorar el servicio de los publicadores.
14. Los datos deben ser auténticos, exponiendo las garantías de calidad y pertinencia de los datos al efecto.

Especial importancia ofrece este último punto, verdadero “talón de Aquiles” del paradigma Open Data. La confianza hacia los datos de aquellos que van a utilizarlos debe estar garantizada, al menos en las publicaciones de datos del sector público

2.3.2 ¿QUÉ CONSECUENCIAS TIENE LA APERTURA DE DATOS?

La apertura de datos permite la reutilización de los mismos generando con ello nuevo conocimiento, nuevas oportunidades económicas, nuevas posibilidades de innovación y potenciales cambios sociales. Open Data es un paradigma que afecta a toda la Sociedad, ciertamente los gobiernos están teniendo un papel fundamental de fomento e impulso en las publicaciones de datos, presentado portales y catálogos de datos cada vez más completos y extendidos, pero también las empresas y los individuos están participando en el fenómeno Open Data aportando su experiencia, ofreciendo nuevas posibilidades de negocio y nuevos modelos de datos compartidos. El sector público no incurre en grandes gastos a la hora de exponer sus datos, éstos se han generado durante su actividad pública y además de los beneficios vertidos a la Sociedad, hay que considerar los conseguidos por la retroalimentación que se obtiene que ordena y optimiza los conjuntos de datos públicos (Buchholtz et al., 2014).

De hecho, la transparencia de las organizaciones, y no sólo las públicas, supone una mejora de su visibilidad, lo que en el contexto corporativo abre una nueva relación de confianza entre las empresas y sus clientes, lo que otorga a las “Empresas Open”, una mejor posición de negocio. Conviene recordar aquí lo referido anteriormente sobre Big Data: el análisis de los macrodatos puede generar beneficios, que se verán aumentados si están en acceso abierto.

La transparencia afecta a la eficiencia de los mercados y a medio plazo puede mejorar la competitividad y generar una comunicación bidireccional que identifique mejor la demanda y adecue la oferta de productos y servicios. También aumenta la posibilidad de control y rendición de cuentas del sector público pero también el empresarial, mostrando por ejemplo las cifras de inversión en I+D, o el montante dedicado a la responsabilidad social corporativa. El Informe McKinsey sobre datos abiertos introduce la expresión “datos líquidos”, aportando la noción de fluidez de los mismos en relación a nuevas y mejoradas posibilidades de colaboración entre los principales actores de la Sociedad: Gobiernos, empresas y ciudadanos (Manyika et al., 2013).

La filosofía “Open” está en relación directa con el conocimiento abierto y consecuentemente con la investigación abierta. En éste ámbito la apertura de datos mejora la comunicabilidad científica, elimina iniciativas redundantes y genera nuevo conocimiento producido por la reutilización de datos. La vinculación de los datos a través de técnicas semánticas ofrece, nuevas posibilidades de recombinación de la información, mejorando la disponibilidad en red de los datos y potenciando su accesibilidad; de hecho la combinación de licencias de uso abiertas y de técnicas Linked Data constituyen el verdadero núcleo de desarrollo de la reutilización de los datos y por ello de la generación de nuevos e innovadores resultados. El concepto básico es la interoperabilidad como principio que fundamenta los beneficios en la apertura de datos, incrementando las posibilidades de combinación de distintos conjuntos de datos que desarrollen más y mejores productos y servicios (Bauer & Kaltenböck, 2012). Subirats (2012) define la interoperabilidad como: “... *datos*

distribuidos, utilizados e intercambiados por instituciones sin la necesidad de centralizar los datos o estandarizar el software. La interoperabilidad se consigue cuando las máquinas entienden el significado de los datos distribuidos y son capaces de procesarlos de manera correcta.

El estudio demosEuropa 2014 advierte que la mera publicación de datos no es suficiente. La previsión de beneficios futuros pasa por la creación de plataformas de apoyo a la creación de datos abiertos en el marco de una política común europea que implante políticas coordinadas en los Estados de la Unión Europea. También hace referencia a los importantes beneficios económicos que conlleva la aplicación de Open Data, con incrementos acumulados de dos dígitos en PIB europeo hasta el año 2020 y de la aplicabilidad directa de los datos abiertos al 10% de la

economía y hasta en un 33% de modo indirecto.

Conviene mencionar, a nivel nacional, los datos aportados por el Estudio de Caracterización del Sector Infomediario, publicado por el Observatorio Nacional de las Telecomunicaciones y la Sociedad de la Información en 2012, que las iniciativas Open Data y de reutilización de la información del sector público (RISP) generaron una cifra de negocio entorno a los 440 millones de euros al año, aportando casi cinco mil empleos al sector de la reutilización de la información. Las empresas infomediarias, dedican aproximadamente la mitad de su capacidad a la intermediación sobre los datos de la Administración.

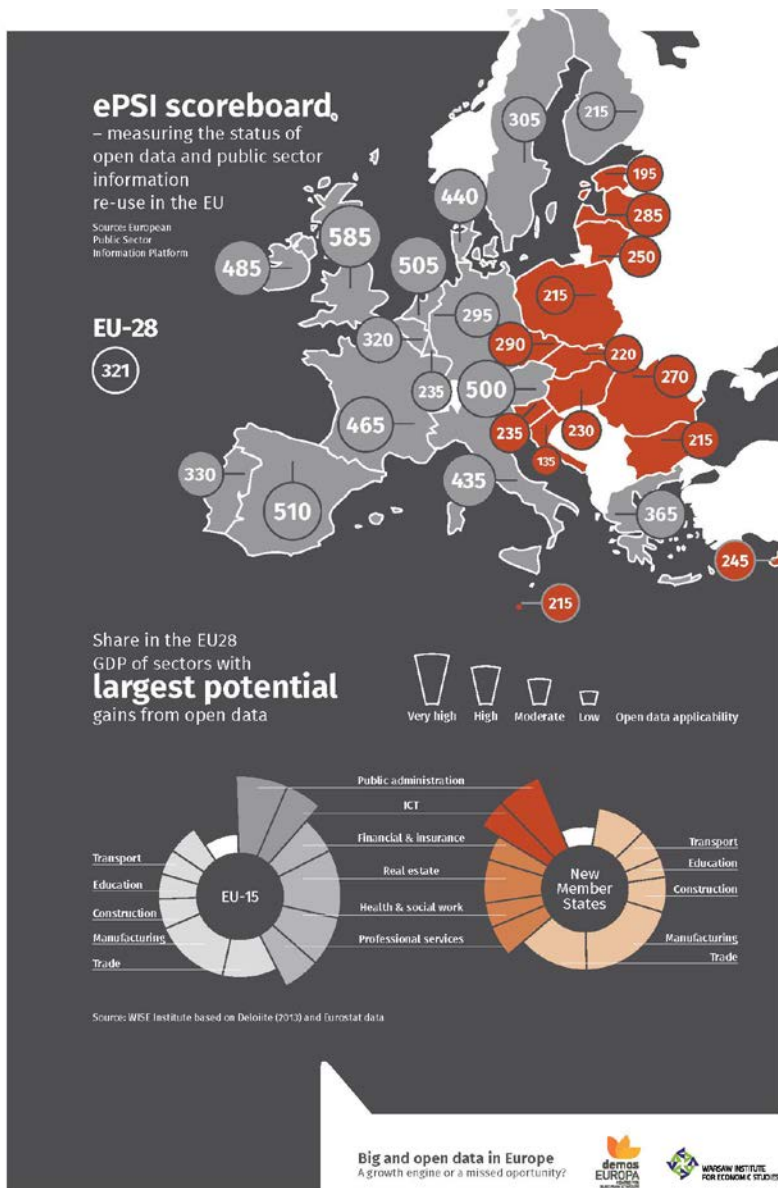


Figura 2 Índice ePSI referenciado al empleo de Open Data y el nivel de reutilización de la información del sector público.
Fuente: demosEuropa 2014

2.3.3 OPEN GOVERNMENT: EL MOTOR DEL CAMBIO

Las iniciativas de Gobierno abierto, como paradigma de la transparencia, la colaboración y la participación de los ciudadanos en la gestión de los organismos públicos, se han mostrado como el verdadero motor de la apertura de datos. Sus orígenes datan del año 2009, tras la declaración de Barack Obama en The Memorandum on Transparency and Open Government. De esta iniciativa deriva el Open Government Data (OGD) como un movimiento político, económico y social, cuya finalidad es abrir los datos de carácter público para la reutilización por la sociedad civil, los agentes económicos, los entes culturales, la clase política, los administradores públicos etc. (Bauer & Kaltenböck, 2012).

En 2011 nace el Open Government Partnership, como plataforma para los administradores públicos que fomenten la apertura de los gobiernos, la rendición de cuentas y la mejora de la respuesta a los ciudadanos. Esta organización cuenta en 2014 con 63 miembros, en cuyos países trabajan las administraciones públicas y la Sociedad, en el desarrollo e implementación de reformas hacia el Gobierno abierto. Las razones que promueven el Gobierno abierto pueden sistematizarse del siguiente modo (Open Government Partnership, 2013):

1. El acceso libre de los ciudadanos a los datos gubernamentales supone recibir información de funcionamiento de las administraciones y por ello de conseguir una imagen más certera sobre la calidad de su actuación. La transparencia obliga a las administraciones a presentar la información veraz y exhaustiva de su actividad, con formatos que permitan compartir y reusar los datos.
2. Liberar el potencial de crecimiento social y económico. La apertura de datos de los gobiernos promueve la innovación y los servicios que aportan valor social y comercial (Open Knowledge Foundation, 2013).
3. Fomento, por parte de los gobiernos, de la participación y cooperación ciudadana. Participación, que se plasme en apoyo activo a las políticas públicas y cooperación, que promueva la actuación conjunta de la Sociedad civil con los entes públicos hacia objetivos innovadores y de mayor efectividad social y política.
4. OGD favorece la rendición de cuentas, al hacer visible la actividad pública. Las administraciones públicas deben justificar su acción y responsabilizarse de los efectos de la misma.

Los resultados económicos no son los únicos presentes, también la posible reconversión social y mejoras en las formas de hacer política. Los valores que aportan las políticas gubernamentales de datos abiertos se pueden sistematizar así (Ubaldi, 2013):

1. Aumento de la confianza en las administraciones públicas: OGD mejora la percepción de las políticas gubernamentales, mientras la transparencia ayuda al propio autocontrol de

los servidores públicos. Este permite una mejor comunión entre ciudadanos y gobernantes.

2. Aumento de la participación social y del compromiso de la población con los gobiernos, alcanzando el auto empoderamiento. El ciudadano se siente parte del proyecto y obtiene beneficios para su propia vida.
3. Mejora de la eficacia, participación y compromiso de los empleados públicos, que adquieren una motivación añadida al formar parte de una organización que se acoge a una democratización de su actuación.
4. Fomento de la innovación, la eficiencia y eficacia en los servicios públicos. Las necesidades de información disminuirían, los costes de actividad serían menores y las peticiones de los gobernados serían más concretas, pudiéndose mejorar, además, los tiempos de resolución de procedimientos.
5. Mejora la calidad de los conjuntos de datos.
6. La economía en su más amplio espectro se vería beneficiada por la actividad derivada del OGD. Nuevos ingresos provenientes de la nueva economía, aplicaciones generadoras de valor, incrementos importantes en el empleo, exportación del modelo de datos al sector privado de la economía.

Es posible preguntarse qué espera la población del Gobierno abierto y si los principales enfoques expuestos cubren esencialmente esas necesidades. El Estudio de la Demanda y Uso de Gobierno abierto en España afirma, en su informe lo siguiente (Márquez Fernández, Vázquez Martínez, Martínez López, & Roldán Cruz, 2013):

1. El 65,5% de la ciudadanía considera como buena o muy buena la calidad de los servicios públicos. Es probable que ese porcentaje sea sustancialmente peor en el año actual (2014) debido a los efectos de la crisis.
2. El 42,2% realizan trámites electrónicamente con las administraciones públicas y el 65,2% consultan las webs públicas para informarse de actividades o actuaciones administrativas.
3. El 75% de los encuestados creen que la transparencia mejoraría la confianza en el Gobierno.
4. El 81,4% de la ciudadanía reclama nuevos canales para incrementar la participación y apuesta por la convivencia entre canales digitales y canales tradicionales.
5. El 53,6% de los usuarios de Internet manifiestan interés en los asuntos políticos frente al 34,8% de los que no lo son.

Se puede asegurar que la ciudadanía está demandando nuevos niveles de democratización gubernamental que pueden ser perfectamente instrumentados por el desarrollo del Gobierno abierto en nuestro país. Open Data es el vehículo adecuado.

2.3.4 MARCO LEGAL DEL OPEN GOVERNMENT DATA

La explicación del marco jurídico que se establece sobre los datos requiere una distinción previa entre el concepto de reutilización de la información y el de Open Data. Éste último considera la información como un bien de acceso público en un contexto de libertad, regida por principios acordados por la comunidad, bajo licencias de acceso libre y gratuito, haciendo hincapié en la adquisición y compartición del conocimiento que permita y favorezca la innovación y la investigación. La Reutilización de la Información del Sector Público (RISP) es un concepto enfocado a la puesta a disposición de la información del sector público con elementos comunes a Open Data, pero con algunas diferencias como la posibilidad de pago por acceso a los datos, regulación de la protección de datos más específica, la apertura de licencias es más selectiva. La orientación final es similar aunque centrada en el fomento del desarrollo económico mediante el aprovechamiento de la información por agentes infomediarios (intermediarios entre los datos públicos y los productos elaborados de datos).

La normativa marco europea se define a través de la Directiva 2003/98/CE, cuya aspiración en el ámbito RISP se orienta a la dinamización del sector privado de la información mediante la generación de nuevos servicios a nivel europeo, basados en los datos del sector público, promoviendo la utilización de esos datos por empresas privadas en todo el territorio de la Unión.

En el año 2009 se produjo otro hito normativo relevante, la Declaración Ministerial de Malmö que propugnaba una mayor participación ciudadana en las políticas públicas y el fomento de acciones favorecedoras de la transparencia.

La Declaración de Granada de 2010 mantiene en sus postulados la importancia del desarrollo de políticas de reutilización de la información.

La Agenda Digital Europea pretende utilizar el empuje de la economía digital para, entre otros objetivos, generar riqueza basada en la reutilización de la información indicando que, el sector público es el elemento dinamizador del mercado de contenidos digitales (Marcos-Martín & Soriano-Maldonado, 2011). La adaptación a nuestro país de los objetivos de la Agenda Digital de Europa se produce a través de la Agenda Digital de España (2013), que se manifiesta del siguiente modo: *“...el desarrollo de una estrategia de gobierno abierto que estimule la transparencia, participación y la colaboración con otros agentes; y la reutilización de la información del sector público para facilitar la generación de valor y conocimiento”*. Finalmente el plan Horizon 2020 insiste en la línea de innovación y progreso en el contexto del Gobierno abierto promoviendo ideas como los servicios públicos personalizados, la persuasión a los jóvenes para que adopten un papel participativo en las actividades del sector público mediante tecnologías reusables, utilizar tecnologías emergentes en la Administración, introducir el *mobile eGovernment*, y la Administración en la nube (European Commission, 2014b).

Este marco legal es modificado por la Directiva 2013/37/UE del Parlamento Europeo y del Consejo, que modifica a la Directiva 2003/98/CE relativa a la reutilización de la información del sector público. Los cambios más relevantes han sido (Ministerio de Industria, Energía y Turismo, 2014b):

1. Derecho a la reutilización de la documentación pública (siguiendo la legislación de acceso interna). Cuestión por otra parte, ya reconocida en el RD 1495/2011.
2. Ampliación del ámbito de aplicación a las bibliotecas, los archivos y los museos públicos, necesitándose para ello la autorización expresa de reutilización, según las normas de la Directiva 2003/98/CE.
3. Se introduce la recomendación de utilizar formatos abiertos, según los estándares reconocidos y legibles por computador.
4. Se recomienda que la reutilización se ofrezca sin condiciones o con las mínimas posibles que deberán estar disponibles en licencias preferiblemente digitales.
5. Las tasas por reutilización se han de limitar a los costes marginales y se deben establecer de modo transparente.
6. Los acuerdos exclusivos para archivos, bibliotecas y museos públicos se permitirán hasta un plazo de diez años.
7. Obligación de rendir cuentas ante las autoridades europeas sobre la implantación de la nueva Directiva.

En cuanto a la legislación española y a falta de transposición de la Directiva PSI 2013, la Ley 37/2007, de 16 de noviembre, sobre Reutilización de la Información en el Sector Público, persigue (Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, 2007):

1. Publicar todos los documentos de libre disposición.
2. Facilitar la creación de productos y servicios de información basados en datos públicos.
3. Fomento del uso transfronterizo de los datos por usuarios e infomediarios.
4. Promover la puesta a disposición de los datos por medios electrónicos.

El Real Decreto 1495/2011, de 24 de octubre, de desarrollo de la Ley 37/2007 cuyos principales contenidos son (Real decreto 1495/2011, de 24 de octubre, por el que se desarrolla la ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal. 2011):

1. Definir el objeto y ámbito de aplicación. (Ahora modificado por la PSI 2013)
2. Desarrollar el régimen jurídico de la reutilización de la información del sector público estatal.
3. Desarrollo del régimen de modalidades de reutilización de documentos.
4. Especificación del régimen a los documentos reutilizables sujetos a derechos de propiedad intelectual o con contenidos personales.

La ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen Gobierno que obliga a publicar los datos sobre temas considerados relevantes (Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno. 2013) .

Y finalmente la Norma Técnica de Interoperabilidad de Reutilización de Recursos de la Información, Resolución de 19 de febrero de 2013, que *“establece las condiciones sobre selección, identificación, descripción, formato, condiciones de uso y puesta a disposición de los documentos y recursos de información elaborados o custodiados por el sector público, relativos a numerosos ámbitos de interés como la información social, económica, jurídica, turística, sobre empresas, educación, etc., cumpliendo con lo establecido en la Ley 37/2007, de 16 de noviembre”*.

Los organismos públicos establecen planes de desarrollo de las políticas de datos para el fomento y la coordinación de actividades. En nuestro país la Agenda Digital para España (Ministerio de Industria, Energía y Turismo, 2014a), establece medidas para promover la publicación de datos. En esta misma línea, el Proyecto Aporta, desde 2009, pretende la mejora del mercado de reutilización de los datos públicos y el apoyo técnico a los organismos públicos, desarrollando la armonización entre organismos, todo ello en el contexto de utilización de los estándares internacionales dispuestos al efecto.

La obtención de la información pública puede hacerse, según el marco jurídico vigente, de tres formas diferentes: acceso completamente libre a la información, acceso mediante solicitud y acceso mediante licencia que fija las condiciones de reutilización. Dada la importancia de este último supuesto y teniendo en cuenta las actualizaciones legislativas previstas, se tratará este tema con más detalle y profundidad en otro epígrafe.

2.3.5 ¿CÓMO ABRIR LOS DATOS?

La Open Knowledge Fundation (2013) nos ofrece una serie de recomendaciones previas a la apertura de datos. En primer lugar se debe abordar un proyecto sencillo que permita obtener resultados ágiles y fáciles de controlar. La apertura debe ser lo más inmediata posible, tras la creación de los datos, ofreciéndolos a la Sociedad y fomentando la utilización por los intermediarios que transformarán la información en aplicaciones más utilizables por la comunidad.

El proceso completo de apertura puede variar de unos conjuntos de datos a otros, pero se pueden definir un proceso básico orientativo.

1. Para la selección del conjunto de datos se ha de tener en cuenta principalmente, las necesidades detectadas entre los usuarios potenciales y el coste de apertura, por ello se

deben publicar datos susceptibles de ser utilizados por su relevancia. Es útil observar qué hacen otros organismos con esta cuestión.

2. Abrir los datos bajo licencia clarifica la utilización por los reutilizadores. Es habitual utilizar una de las siguientes:
 - a. Public Domain Dedication and License Incluye nuestros datos en el Dominio Público sin más restricciones.
 - b. Open Database License que permite que los utilizadores distribuyan los datos o sus productos con una licencia igual a la que regula los datos originales. Esta licencia puede incluir la obligación de atribuir la propiedad de los datos o no.
3. El siguiente paso es la puesta a disposición de los usuarios de los datos. El formato requerido debe ser un estándar que pueda ser procesado fácilmente por computadoras, es el caso de Resource Description Framework (RDF), el estándar más utilizado para la descripción semántica y la vinculación de datos abiertos. Se debe tender hacia la absoluta gratuidad de las descargas de datos y hacia la publicación de los datos en conjuntos completos. La publicación por defecto será en Internet, en una web que recoja los *datasets*, pero existen otras opciones: repositorios de datos, servidores FTP, o incluso una API (*Application Programming Interface*) que permita ya obtener el conjunto completo de datos o partes definidas por el reutilizador. En ningún caso la utilización de interfaces de programación puede suplir a la publicación del conjunto completo de datos en su formato original.
4. Finalmente se podría concluir el ciclo de apertura de datos asegurando su visibilidad, no sólo presente sino también futura. Para ello existen herramientas web que nos permiten publicar nuestros datos en conjunto con otros aumentando la concentración y por ello la visibilidad. El Open Knowledge Foundation indica algunas consideraciones importantes al efecto para agencias gubernamentales: utilizar herramientas de código abierto en vez de generar aplicaciones para catalogar conjuntos de datos propias, permitir que la iniciativa privada publique sus datos en los repositorios de datos abiertos públicos (mejorando la idea de cooperación y participación) y evitar niveles de acceso privilegiados a ciertos sectores de la Sociedad (Open Knowledge Foundation, 2013).

En el contexto RISP de la publicación de datos, la Comunidad Open Data – RISP España, definió un decálogo de buenas prácticas, generado en las conclusiones del Día Open Data en Euskadi, celebrado en San Sebastián en 2012. Este decálogo ha cobrado gran relevancia desde entonces y es considerado un estándar, en el ámbito del proyecto Aporta, para las publicaciones de datos en las administraciones públicas. Se enumera brevemente:

0. Armonización entre Administraciones. (Este punto especial del catálogo pretende definir la prioridad en cuanto a la reutilización de la información: la coordinación de los organismos del sector público que deben compartir los mismos principios en aras de la interoperabilidad y de la importancia del trabajo conjunto)

1. Publicar datos en formatos abiertos y estándares.
2. Usar esquemas y vocabularios consensuados y utilizar metadatos abiertos.
3. Inventario en un catálogo de datos estructurado.
4. Datos accesibles desde direcciones web persistentes y amigables.
5. Exponer un mínimo conjunto de datos relativos al nivel de competencias del organismo y su estrategia de exposición de datos
6. Compromiso de servicio, actualización y calidad del dato, manteniendo un canal eficiente de retroalimentación entre el reutilizador y las administraciones públicas.
7. Monitorizar y evaluar el uso y servicio mediante métricas.
8. Datos bajo condiciones de uso no restrictivas y comunes.
9. Educar en el uso de datos.
10. Recopilar aplicaciones, herramientas y manuales para motivar y facilitar la reutilización.

2.3.6 LOS PROBLEMAS DE LA APERTURA DE DATOS

Open Data y la reutilización de la información pública tienen algunos problemas que resolver. Parece claro que la información personal debe quedar protegida, la información afectante a la Seguridad Nacional queda también al margen de la apertura de datos, del mismo modo que los datos afectados por cuestiones de propiedad intelectual. Pero en un ecosistema de datos abiertos y probablemente vinculados, la total protección es difícil de conseguir, las sucesivas re combinaciones de datos puede llegar a mostrar información sensible. Es complejo que los propietarios de los datos sepan prever siempre las consecuencias de la publicación. Además, una vez abiertos, las posibilidades de cerrar con posterioridad los datos son escasas. Es por ello que se debe usar lo que algunos autores llaman la libertad negativa, es decir, el establecimiento de reglas claras de lo que se debe o no se debe abrir. Aun así es posible acceder a la información de esos datos, siempre y cuando se pueda disgregar la información de los datos, de aquella parte que los personaliza o vincula con la información sensible (Buchholtz et al., 2014).

Las empresas pueden aprovechar la apertura de sus propios datos para generar nuevas posibilidades de negocio y apoyar la investigación. La otra cara de la moneda en este punto es el aprovechamiento desleal de esos datos por la competencia, que evidentemente puede suponer un cierre de la información. La transparencia hacia el cliente también puede comprometer la

imagen de la empresa permitiendo crisis de reputación más severas si no se considera por el consumidor la faceta positiva de la apertura de la información (Manyika et al., 2013).

Los costes asociados a la apertura pueden pesar negativamente en las cuentas de las empresas, a diferencia del sector público, no está tan claro que la retroalimentación compense el esfuerzo, al menos en el corto plazo. La mera apertura de datos puede no ser suficiente, se ha de exigir a los intermediarios la generación de herramientas que aporten valor y con políticas bien pensadas sobre propiedad intelectual o la privacidad y la protección de los datos sensibles, pues en otro caso se corre el riesgo de un retroceso en el movimiento de apertura de datos. La inversión en análisis de datos, catalogación y limpieza de datos debe potenciarse. Para permitir la extracción de conocimiento, se debe mejorar la descripción de los conjuntos de datos con metadatos que los identifiquen correctamente mejorando su usabilidad, se deben crear procesos de protección de actuaciones mal intencionadas sobre los datos y desarrollar y formar el talento necesario para que esas herramientas puedan ser utilizadas, la carencia de “profesionales de los datos” es uno de los problemas más importantes a solucionar (Manyika et al., 2013).

Según Kitchin, existen cuatro debilidades que rodean el entorno Open data. En primer lugar se está priorizando la publicación, sin embargo no están evolucionando en la misma medida las programaciones económicas y técnicas para hacer estos esfuerzos publicadores sostenibles. En segundo lugar, avisa del peligro de que la democratización y transparencia, que posibilitan los datos abiertos, no se queden en meras declaraciones de intenciones, vigilando que los datos públicos no sirvan únicamente para generar negocio en el mercado infomediario, mientras que pueden existir sectores de la ciudadanía que no acceden por el coste a los beneficios de las aplicaciones de datos, ni a la información gubernamental. El tercer punto se refiere a la verdadera usabilidad y utilidad de los datos abiertos; en no pocas ocasiones los datos no están bien organizados, ni son de calidad, ni cumplen con los requisitos de interoperabilidad, ni contienen metadatos o no son adecuados, tampoco se contempla ningún plan de preservación, copia de seguridad, ni política alguna de auditoría. Finalmente, la cuarta consideración del autor se refiere a la vinculación política de los datos. Open Data debe por definición estar alejado de ideologías y posturas políticas o económicas determinadas, pero en ocasiones no es así. Expone Kitchin el caso del Reino Unido, donde algunas publicaciones de datos sustentan políticas de austeridad y estrategias de privatización de los servicios públicos.

2.4 DATOS ABIERTOS DE INVESTIGACIÓN

2.4.1 UNA APROXIMACIÓN AL TEMA

Podemos definir los datos de investigación, (National Institutes of Health & Organización para la Cooperación y Desarrollo Económico) como aquellos datos registrados durante la misma, que se utilizan como fuentes primarias para la investigación científica, con reconocimiento de la comunidad científica y que verifican los resultados de la investigación. Los elementos que no se consideran datos de investigación son los cuadernos de laboratorio, proyectos de artículos científicos, planes de futuras investigaciones, revisiones de pares o las comunicaciones con colegas (OECD, 2007; Torres-Salinas, Robinson-García, & Cabezas Clavijo, 2012).

La tipología de datos de investigación es muy variada, de ahí la complejidad de su gestión. Sin pretender ser exhaustivos podemos afirmar que datos de investigación son: las medidas de los instrumentos científicos, los datos derivados de la observación experimental, las imágenes, el video y audio producidos en el contexto de la investigación, documentos de texto, hojas de cálculo, bases de datos, datos cuantitativos, resultados de las encuestas, transcripciones de entrevistas, modelos de simulación y software, presentaciones, ejemplos, etc. (Jones, Marieke, & Pickton, 2013).

La comunidad científica está a favor de la puesta en común de los datos de investigación y la comunidad documental ya está definiendo el marco de trabajo para cristalizar este movimiento. Los retos son importantes, pues el manejo de los datos requiere técnicas de gestión de datos complejas: políticas de protección de la privacidad, archivo y presentación, sistemas de recuperación de los datos, creación de procesos de preservación y de accesibilidad y la creación de plataformas que permitan el almacenamiento (Torres-Salinas, 2010).

Aunque el sentimiento general es favorable a la apertura de datos, existen ciertas reticencias en el colectivo investigador a la hora de compartir sus datos, pues existe el temor de que se ponga en entredicho su investigación o que quizás puedan perder cierto reconocimiento. Otros puntos en contra de la publicación abierta de los datos son su utilización fraudulenta y las rigideces que establecen los sistemas de protección de los mismos. (Torres-Salinas et al., 2012).

2.4.2 DATA SHARING

Compartir los datos de investigación tiene importantes beneficios para la comunidad científica y para la sociedad en la que se integran. Podemos enumerar algunos de ellos:

1. Mejor aprovechamiento de los recursos invertidos en investigación.
2. Evitar trabajos duplicados.

3. Mejorar la transparencia evitando el fraude (pues permitiría replicar las investigaciones de modo más sencillo).
4. Aumento de citas (trabajos + citas).
5. Extensión de la filosofía “Open” entre la comunidad científica.

A estos efectos la declaración de Denton sobre Open Data en el contexto de los datos científicos aporta un marco resumen de los beneficios del *Data sharing*. Se trata de un compendio de las mejores prácticas y tendencias en la gestión de datos, definidas por los actores más relevantes del tema: investigadores, bibliotecarios, personal técnico, y gestores universitarios.

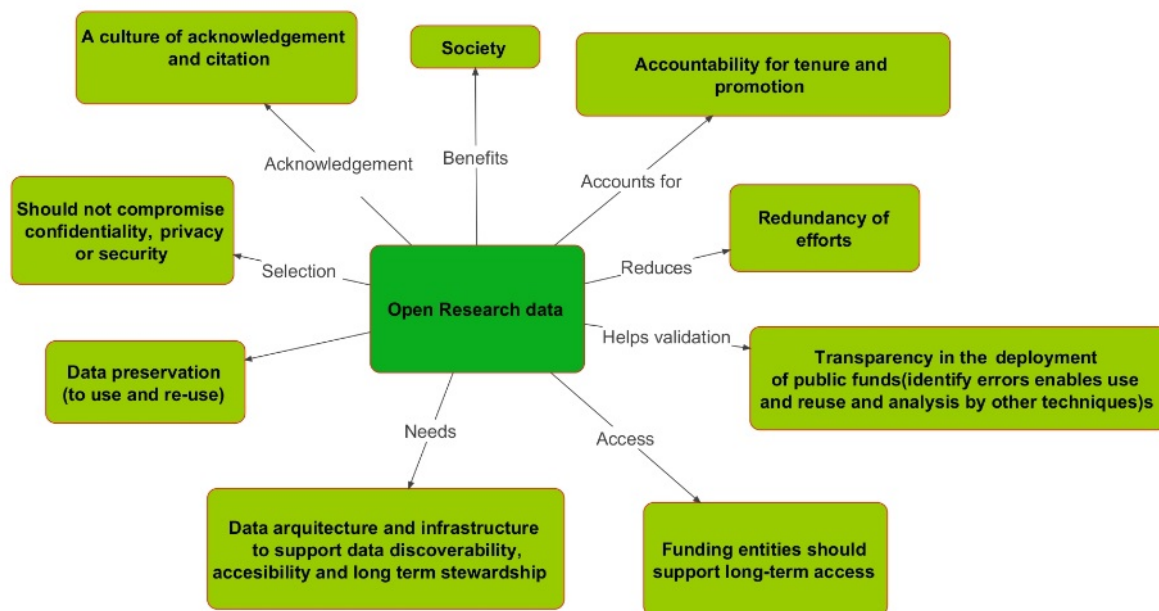


Figura 3 Universo de los datos de investigación. Fuente: Declaración de Denton (2012)

Sus principales directrices hacen mención a la importancia del acceso abierto a los datos de investigación como pieza fundamental del avance de la ciencia, la erudición y la sociedad. Proponen que la investigación financiada públicamente debe estar obligatoriamente en abierto; que la transparencia de la investigación es fundamental para mantener la confianza de la sociedad en la actividad investigadora; que la validación por pares de los datos de investigación es una función necesaria y responsable de la actividad investigadora, y que la gestión integral de datos de investigación es responsabilidad de toda la comunidad a la que afecta (University of North Texas, 2014) .

2.4.2.1 Repositorios de datos de investigación

Es en el entorno universitario donde las infraestructuras para el tratamiento global de los datos tienen un sentido principal. Ya en repositorios independientes, ya en el mismo repositorio que los trabajos de investigación, los datos tendrán de modo necesario, que ocupar el lugar correspondiente junto a las publicaciones de trabajos de investigación, a las que aportan la validez de sus resultados. La ventaja de los repositorios independientes se aprecia en la mejor especialización, contando con características más adaptadas al objeto de almacenamiento. Una importante parte de la comunidad científica busca una mayor personalización a través de los repositorios temáticos que se adaptan mejor a cada una de las disciplinas. También se cuenta con repositorios multidisciplinares especializados en datos por ejemplo Dryad, Figshare, Zenodo o Dataverse (Nina-Alcocer, Blasco-Gil, & Peset, 2013).

Sin embargo los repositorios institucionales existentes ofrecen una mayor experiencia en gestión, almacenamiento, publicación y preservación, aunque con estructuras afinadas para publicaciones científicas en su ámbito. Los repositorios institucionales cuentan con la ventaja de ser interoperables, contar con estructuras que responden de la calidad y continuidad del servicio, y disponer de personal capacitado para tratar con objetos digitales (Hernández-Pérez & García-Moreno, 2013). Existe una iniciativa española que ayuda a los científicos a la hora de decidir dónde publicar sus datos, se trata del proyecto ODISEA que se configura como un directorio de depósitos de datos de investigación a nivel mundial.

2.4.2.2 Cuestiones legales.

La apertura de los datos, cualquiera que sea su tipología ha de ser respetuosa con la legislación sobre propiedad intelectual y protección de datos. La exigencia de confidencialidad y privacidad debe ser respetada en cualquier caso, evitando que datos personales sean compartidos sin permiso. La disponibilidad de los datos puede requerir el consentimiento informado que permita tanto el consumo actual como futuro. Respecto a la confidencialidad, puede ser necesario despersonalizar los datos de tal modo que impidan la identificación de los sujetos afectados y también regular el acceso, definiendo los requisitos para acceder y su régimen de utilización. Se debe asignar una licencia de utilización que informe de las posibilidades de uso futuras. (Jones et al., 2013).

Cuando la investigación utiliza datos de fuentes ajenas, las consideraciones de propiedad intelectual deben ser establecidas al comienzo de la investigación, debiendo identificar la disponibilidad de uso respetando las indicaciones de reutilización. La concreta interpretación de la licencia de utilización debe marcar el futuro uso de los datos.

Como ahora se verá, no existe un marco legal exactamente dirigido a la regulación de los datos de investigación en la UE. En el estudio sobre protección, acceso y uso de los datos de investigación, llevado a cabo por la Universidad de Gotinga (Alemania), se efectúa un análisis del marco legal y sus debilidades. Sus principales consideraciones son las siguientes:

1. Los datos de investigación en sí mismos considerados no están protegidos por leyes de copyright. Sólo las bases de datos y sus estructuras están protegidas (Directiva 96/9/CE del Parlamento Europeo y del Consejo, de 11 de marzo de 1996, sobre la protección jurídica de las bases de datos).
2. La Directiva no cubre la utilización masiva de datos para análisis científicos.
3. La no limitación de uso de las bases de datos para usos científicos es una disposición opcional de la Directiva y no está armonizada coherentemente en los Estados de la UE.
4. La reproducción completa de la bases de datos desde el punto de vista digital, no está permitida según la Directiva mencionada.
5. El uso científico indirecto de las bases de datos no está cubierto por las limitaciones de las Directiva.
6. La vinculación de datos y las publicaciones científicas no están recogidas ni en la Directiva de bases de datos ni en la Directiva de la Sociedad de la Información (Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society)
7. Respecto a la licencias de uso de datos, las licencias Creative Commons 3.0 no abarcan los derechos de bases de datos. La licencias Open Data Commons tienen dificultades de interoperabilidad, no cubren los fines científicos específicos, ni los derechos de autor o bases de datos, por ello para una protección completa se requiere el uso de más de una licencia (Universidad de Gotinga, 2013).

2.4.2.3 Formatos, interoperabilidad y Linked Data

Respecto a los formatos que se vayan a utilizar, lo mejor es establecer formatos abiertos, no propietarios, que se ajusten en lo posible a los estándares internacionales y con posibilidad de estructuración de los datos. La utilización de técnicas de vinculación de datos en la interconexión de los metadatos descriptivos de los datasets producto de la investigación, posibilita una mejora de la interoperabilidad y mayores posibilidades de descubrimiento y por ende de utilización.

Las tecnologías Linked Data ofrecen la posibilidad de una mayor integración de los datos de investigación. La interoperabilidad entre los conjuntos de datos se mejora y a la vez permite técnicas de análisis de estos datos de modo automático. Los Internationalized Resource Identifiers (IRI) permiten una identificación eficaz, incrementando los niveles de recuperabilidad y el acceso, añadiendo además posibilidades de interconexión con otros datos relacionados y

relevantes, lo que puede llegar a generar una red de datos de gran valor. También, la recolección automática de datos y metadatos es más efectiva y no sólo a nivel de datos sino también de las propias publicaciones (The Royal Society, 2012).

Linked Data permite la interconexión de bases de datos, aportando descripciones de metadatos que en su integración permitirán inferir conocimiento de esos datos. De ahí la importancia de la estandarización de la descripción de datasets y la conveniencia de la normalización de identificadores, lenguajes descriptivos, ontologías y vocabularios. De no evolucionar hacia en ese camino, la no normalización puede impedir una interoperabilidad consistente de los datos, si los vocabularios, por ejemplo, carecen de la suficiente homogeneidad, las descripciones se podrán interconectar, pero no se podrá extraer nuevo conocimiento de ellas. Otro problema relevante es el de la obsolescencia de los datos, las tecnologías de vinculación no garantizan la actualización de los set de datos en red y esto puede afectar de modo fundamental a su calidad y a su uso científico.

De importancia básica es que Linked Data vincula conjuntos de datos científicos aportando información sobre su procedencia y los flujos de trabajo empleados. RDF y OWL, ofrecen la posibilidad de deducir conocimiento de los set de datos; pero carecen de capacidad para describir datos fundamentales para la investigación y que son necesarios para dar soporte al proceso, como la descripción del ciclo de vida de los datos y su estado, las cuestiones relacionadas con la propiedad de los datos y el control de versiones que es fundamental para mantener su calidad de a través de actualizaciones y las citas de los datasets.

Se propone una capa de agregación de metadatos que se sitúe por encima de la capa Linked Data. Es un sistema para describir la agregación de los recursos, los *research objet* (RO). Esto permite referir como los datos contribuyen a la investigación, las relaciones que se establecen entre los datos, descripciones del valor adicional de los recursos en conjunto y todo ello en un solo objeto digital fácilmente reusable e intercambiable (Bechhofer et al., 2013).

Los trabajos de investigación, pueden aprovechar las tecnologías de vinculado para establecer redes de conocimiento. Los resultados de la investigación pueden ofrecerse a través de nuevos productos como son los *executable paper* que permiten la interacción a través de vínculos de la información de investigación tanto estática como dinámica y permite la navegación por los datos, los resultados y otras partes interactivas o no del documento bajo soporte web (The Royal Society, 2012).

2.4.2.4 Citación de los datos abiertos de investigación

Como se ha dicho anteriormente, abrir los datos para el aprovechamiento de otros científicos supone un aumento de citas a sus productores. Existen estudios que aseveran que el aumento de citas se produce en porcentajes superiores al 30%. Efectivamente, los conjuntos de datos son una

unidad de información en sí misma considerada y si su valor es apreciable, la cita se ha de producir directamente sobre el set de datos. El mismo sistema que se deduce de las citas recibidas por revistas científicas y otros trabajos de investigación, es aplicable como medida de calidad y de impacto en la ciencia, respecto a datasets muy citados. Tradicionalmente los sistemas de cita científica están ya muy asentados, pero no tanto la estructura de citación de los datos. Existen iniciativas que persiguen la normalización de estos sistemas, ofreciendo pautas y buenas prácticas al efecto, así es el caso del proyecto SageCite que nos indica una serie de premisas para las citas de los conjuntos de datos (Ball & Duke, 2012):

1. La cita ha de identificar de forma unívoca y exclusiva el dataset citado.
2. La cita debe describir tanto el conjunto completo de datos como sus subconjuntos.
3. Debe enlazar a través de “HTTP” con la infraestructura que contiene los datos.
4. Debe ser legible por máquina, lo que permita asociar servicios automáticos como las métricas de datos, servicios de descubrimiento y enlace etc.

Las grandes empresas de contenidos científicos, están posicionándose a la hora de facilitar la visibilidad y la posibilidad de citar los conjuntos de datos. Es el caso de Thomson Reuters, que hace poco más de un año completó su portal de la ciencia con otro servicio más, el Data Citation Index que pretende ser punto de acceso a los datasets en un amplio abanico de áreas de conocimiento, aunque más del 80% de los conjuntos de datos responden a investigación en el área de las ciencias. El tipo de registro que indiza fundamentalmente es el propio dataset y los estudios de datos (Torres-Salinas, Martín-Martín, & Fuente-Gutiérrez, 2014).

2.4.3 EL PROGRAMA HORIZON 2020

Horizon 2020 es el programa para la innovación y la investigación, promovido por la Unión Europea para el período comprendido entre los años 2014 y el 2020. Dentro de este macro programa se establece un proyecto piloto sobre Open Research Data, que pretende que los investigadores, cuyos proyectos estén financiados por la UE, compartan los datos de investigación para su uso por la comunidad científica, las empresas innovadoras y la ciudadanía. Este programa piloto persigue una mayor eficiencia en los usos científicos y un progreso de la transparencia hacia la sociedad de la actividad investigadora, buscando facilitar el acceso a esa información y permitiendo la reutilización, lo que redundará sin duda, en mejoras en la innovación, la investigación y la economía.

Siete áreas de conocimiento son puestas a prueba con el programa piloto, cuyas actividades serán apoyadas con más de 3000 millones de euros; estas áreas son las siguientes (European Commission, 2014a):

1. Tecnologías de futuro y emergentes.
2. Infraestructuras de investigación, dentro del marco e-infraestructuras.

3. Liderazgo en tecnologías industriales, de información y de comunicación.
4. Desafío social: Energía eficiente, segura y limpia, dentro del programa Ciudades y Comunidades inteligentes.
5. Desafío social: Acción por el clima, medio ambiente, eficiencia de los recursos y materias primas, exceptuando las líneas relacionadas con las materias primas.
6. Desafío social: Europa en un mundo cambiante, sociedades inclusivas, innovadoras y reflexivas.
7. Ciencia con y para la Sociedad.

El *Model Grant Agreement* (modelo de subvenciones del Plan Horizon 2020), hace referencia normativa a los datos de investigación en el artículo 29.3 (European Commission, 2013); en él se indica (siempre dentro de las condiciones del Plan Piloto) que los participantes deben depositar en un repositorio de datos el producto de su investigación, permitiendo a terceras partes el acceso, la minería de datos, la explotación, reproducción y difusión de los mismos de modo gratuito, debiendo incluir por un lado, los datos necesarios para validar los resultados presentados en las publicaciones científicas, junto con los metadatos descriptivos de los mismos y otros datos, incluidos también sus metadatos asociados, afectantes a la investigación, dentro de los plazos establecidos en el plan de gestión de datos.

2.4.3.1 Plan de gestión de datos (DPM)

El Plan de Gestión de Datos (DMP) es un documento de carácter obligatorio para aquellos proyectos de investigación que se desarrollen y se financien dentro del Open Research Data Pilot. No es lugar aquí para hacer un profundo desglose de un plan de gestión de datos, si remarcar que no es un documento de estructura completamente definida, pues evoluciona a lo largo del proyecto y es flexible al contexto de investigación en el que nos encontremos, aunque está sujeto a unos requisitos de contenidos mínimos (European Commission, 2013):

1. Nominar el conjunto de datos de referencia del proyecto.
2. Describir el conjunto de datos capturados o generados: su procedencia, naturaleza y utilidad y si apoyan una publicación científica. Se incluirá información sobre conjuntos de datos similares y las posibilidades de integración y reutilización.
3. Utilización de metadatos según los estándares establecidos y si no existen en la disciplina, describir los que se crearán.
4. Descripción del proceso de compartición de datos y los métodos de acceso; definición de los períodos de embargo si los hubiera, mecanismos técnicos necesarios para la difusión, software y herramientas necesarios para la reutilización de los datos. Indicación del nivel de apertura de los datos y del repositorio que los contiene indicando el tipo. Se debe justificar la no compartición de los datos.

5. Descripción de los métodos de preservación de los datos, indicando el tiempo que deben conservarse los datos y los costes que esto supondrá y cómo se conseguirán.

2.4.4 CURACIÓN DE DATOS DE INVESTIGACIÓN

Como se ha referido anteriormente, los datos de investigación requieren un esfuerzo de gestión previo a su creación. Existen muchos actores involucrados en la cuestión: científicos, editores, informáticos, gestores institucionales, profesionales de la información, etc., todos ellos involucrados en esfuerzos conjuntos hacia una gestión de estructura compleja cuyo producto de salida es el conjunto de datos de investigación definido, abierto y preservado. De modo genérico se podría decir que toda actividad de gestión, a estos efectos puede ser considerada curación de datos (Martinez-Urbe & Macdonald, 2008).

La curación digital comprende la gestión, el mantenimiento, la preservación y el aporte de valor a los datos de investigación a lo largo de su ciclo de vida. Esta actividad conserva el valor de los datos, minimiza la obsolescencia digital, evita la duplicación de esfuerzos en la creación de datos de investigación y genera valor al compartir los datos para otras actividades de investigación (Digital Curation Centre, 2014; RECOLECTA. Grupo de Trabajo de Depósito y Gestión de datos en Acceso Abierto, 2012).

La curación de datos de investigación participa en todas las fases del ciclo de vida de los datos, procurando el aprovechamiento y la reutilización y el aumento de valor. Digital Curation Centre nos los esquematiza del siguiente modo (Digital Curation Centre, 2014):

1. Datos de investigación como materia prima: ya objetos digitales simples o complejos o datos derivados de registros estructurados.
2. Fases estructurales del ciclo de vida de los datos:
 - a. Describir y representar la información: asignación de metadatos.
 - b. Plan global de preservación de datos durante todo el ciclo vital de los mismos.
 - c. Participación en la comunidad, manteniendo la vigilancia sobre los cambios que se produzcan en el contexto de la gestión de datos de investigación, manteniendo una posición activa en cuanto a normas y herramientas.
3. Curación de datos de investigación propiamente dicha:
 - a. Planificar la creación o el uso de los datos.
 - b. Crear datos (o recibirlos) y sus metadatos descriptivos, administrativos, estructurales y de preservación.

- c. Evaluar y seleccionar los datos teniendo en cuenta la conservación a largo plazo y las tareas de curación de datos.
 - d. Transferencia, archivo o depósito de los datos, adecuándose a las políticas definidas.
 - e. Preservar los datos, asegurando que mantienen su autenticidad, fiabilidad, utilidad e integridad; para ello se establecen acciones de limpieza de datos, comprobación de validez, asignación de metadatos y se asegura la representación de datos y el mantenimiento de formatos.
 - f. Almacenamiento de datos de forma segura.
 - g. Mantenimiento a largo plazo de la posibilidad de acceso y reutilización.
 - h. Transformación de los datos en nuevos subgrupos que permitan su reutilización o que manifiesten un mayor valor.
4. Acciones optativas:
- a. Expurgo de los datos no seleccionados.
 - b. Reevaluación y control continuo de los conjuntos de datos.
 - c. Migración de datos a formatos de más fácil preservación en el futuro.

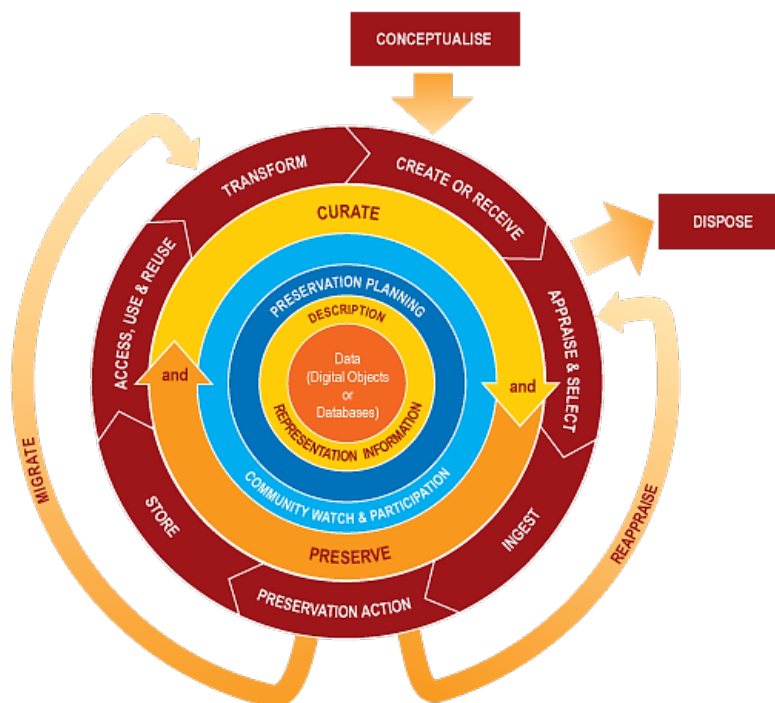


Figura 4 Esquema del proceso de curación de datos de investigación. Fuente: Digital Curation Centre (2013)

2.4.5 LOS ACTORES DE LA GESTIÓN DE DATOS. PROFESIONALES DE LA INFORMACIÓN E INVESTIGADORES

2.4.5.1 Los profesionales de la información

Las bibliotecas y en sentido más amplio, el sector profesional de la información, han encontrado un nuevo foco de contenido donde desarrollar sus competencias. Los investigadores esperan apoyo y asistencia técnica por parte de los servicios de información, sobre todo en las áreas de preservación, documentación, asesoría legal, gestión de repositorios y apoyo a la publicación y descripción con vocabularios controlados (RECOLECTA. Grupo de Trabajo de Depósito y Gestión de datos en Acceso Abierto, 2012).

Los servicios de información deben implicarse de un modo más intenso y efectivo en la gestión de los datos de investigación. El profesional de la información puede aportar competencias clave para ayudar en la selección de datos, ayudando a los investigadores en la valoración previa del coste de preservación, el coste de regeneración de los datos, el beneficio esperado de la reutilización de los datos y el registro de datos. Si el investigador posee su propio centro de datos se puede mejorar la gestión mediante la adición o perfeccionamiento de los metadatos descriptivos de los datasets. También se puede ejercer un papel central en la auditoria de datos, sobre todo relacionada con la propiedad de los mismos a través del tiempo.

La experiencia en acceso abierto permite al bibliotecario asesorar sobre la publicación de los datos. Una mayor accesibilidad proporciona ventajas evidentes como la justificación y refutación de las investigaciones, mayor transparencia y fiabilidad, mejoras en la recuperación de la información expuesta en datos, mayor cantidad de citas de las publicaciones matriz de los datos, mayor reutilización, posibilitando la reformulación de las investigaciones mediante vinculación de datasets a través de técnicas de vinculación de metadatos descriptivos.

Los retos y dificultades son importantes, pues el experto en información no domina el área de conocimiento de las investigaciones, aunque esto no debería dificultar su participación de modo general en la gestión de datos, ni en el empleo de su experiencia en el trabajo en equipos multidisciplinarios, como es el caso. Las lagunas competenciales se pueden complementar con la experiencia de trabajo y la formación.

La Universidad también se beneficia internamente. En el contexto docente, los datos pueden ser utilizados como complemento idóneo de la formación científica, mejorando la comprensión respecto a los ejemplos teóricos (Ashley, 2012; Torres-Salinas, 2010).

El esfuerzo en formación del personal de los servicios de información tiene que ser importante. Existen carencias fundamentales en las competencias necesarias que conforman el perfil del bibliotecario de datos que deben ser subsanadas: competencias estadísticas, herramientas de representación de la información, herramientas de análisis de datos, etc.

2.4.5.2 Productores de datos: investigadores

Investigadores y científicos presentan unas consideraciones diferentes. En general son quienes generan los datos, pero también pueden ser usuarios de los mismos al reutilizar datos de otras investigaciones.

El contexto del personal de investigación es complejo y sujeto a muy diferentes influencias. En septiembre del 2013, DIGITAL CSIC, efectúa una encuesta entre su comunidad científica a cerca de la gestión de sus datos científicos, su disposición a compartirlos y las necesidades de apoyo exterior. Es cierto que sería arriesgado extrapolar los resultados a toda la comunidad científica española, pero sí puede hacernos ver un posible marco general de la situación en la que nos encontramos (CSIC, 2012; RECOLECTA. Grupo de Trabajo de Depósito y Gestión de datos en Acceso Abierto, 2012).

Los resultados de dicha encuesta resaltan que queda mucho camino por hacer y que se requerirán esfuerzos desde todas las partes interesadas para conseguir resultados efectivos. Se aportan los resultados clave dado su interés:

1. Los planes de gestión no se realizan de modo completo ni sujetos a estándares internacionales. Los científicos e investigadores no lo ven como un punto central en sus flujos de trabajo.
2. No se tienen definidas claramente las diferentes categorías de datos científicos.
3. La gestión de datos se suele realizar de modo personal utilizando medios locales, sin apenas estrategias para recuperar esa información, la despersonalización de los datos, o la protección de la confidencialidad de los mismos.
4. La descripción con metadatos de los conjuntos de datos científicos no es muy frecuente.
5. El acceso en abierto tampoco es considerado una prioridad. Estas son las causas:
 - a. Temor a perder la autoría.
 - b. Desconocimiento de la situación legal de los datos compartidos.
 - c. Desconocimiento de los modos de apertura y los servicios que lo ofrecen.
6. No se percibe el apoyo institucional, falta promoción e información sobre estos temas.
7. No son claras las políticas de las agencias financiadoras en cuanto a datos de investigación.
8. La comunidad investigadora no sabe qué apoyo puede recibir de las instituciones. No se tiene claro el papel de los profesionales de la información ni por parte de la comunidad científica, ni por los propios profesionales.
9. La gestión de la preservación no sigue estándares definidos, ni se siguen procesos mínimamente profesionales a la hora de conservar los datos.

Los investigadores deben percibir mejoras en sus procesos de investigación para involucrarse de modo más intenso y efectivo en la gestión de datos, su apertura y conservación. Estas mejoras son tangibles si se establece un proceso profesional de gestión que permita menores esfuerzos en los procesos de datos, trabajando en un entorno definido y estructurado, con soporte de las instituciones de investigación. Por otro lado las posibilidades de reconocimiento aumentan con la publicación de datos de calidad, ya se ha hablado del aumento de citaciones, pero también hay que considerar las mejoras en reputación que provoca la transparencia de los procesos de investigación. También están influyendo los requisitos legales y normativos que poco a poco se van imponiendo y que “obligan” a la gestión efectiva de los datos (Horizon 2020). (Sigit-Sayogo & Pardo, 2013) .

2.5 ¿QUÉ ES LINKED DATA?

Linked Data es la solución tecnológica que nos ofrece todo un sistema de buenas prácticas para la publicación de datos y su interconexión, permitiendo la interoperabilidad, la reutilización y utilizando como vehículo estándares establecidos por la comunidad LD. Linked Data convierte los datos en registros de información, generando una base de datos global bajo tecnologías web y describe de modo estructurado los datos o conjuntos de ellos haciéndolos accesibles tanto para la comprensión humana, como para los sistemas informáticos (Bauer & Kaltenböck, 2012; Berners-Lee, 2009; Bizer, Heath, & Berners-Lee, 2009). Linked Data es el sustrato de la Web de datos, que se constituye como una red global interconectada a través de enlaces cualificados y semánticos, que unen nodos de conocimiento y fundamentalmente orientada al proceso automático y secundariamente por el ser humano.

Linked Data facilita el acceso a los datos y su descubrimiento mediante tecnologías semánticas de búsqueda, promoviendo la utilización de conjuntos de datos para proveer servicios y aplicaciones novedosas. Para ello, Linked Data utiliza las tecnologías web como soporte, aprovechando la madurez de sus estructuras, la ubicuidad de la Web y su naturaleza distribuida y escalable (Bizer et al., 2009; Heath & Bizer, 2011).

Guerrini y Possemato nos ofrecen otra definición complementaria: “...a los datos publicados de modo legible, interpretable y utilizable por un sistema informático, cuyo significado se define explícitamente por una cadena de palabras y marcadores, constituyéndose en una red de datos vinculados pertenecientes a un dominio, conectado con otros conjuntos de datos externos en un contexto de relaciones cada vez más extendido” (Guerrini & Possemato, 2013).

2.5.1 UNA PERSPECTIVA GENERAL

2.5.1.1 *Los cuatro principios Linked Data*

Berners Lee (2009) introdujo en su declaración sobre Linked Data de 2006, los principios básicos de la vinculación de datos. Dichos principios están orientados a la publicación e interconexión con el objetivo de conseguir la interoperabilidad de los datos y aprovechando tanto la arquitectura como los estándares Web:

1. Usar URI (Uniform Resource Identifiers) para identificar las cosas.
2. Usar URI HTTP, para que las personas puedan buscar esas cosas.
3. Cuando alguien busca una URI, suministrar información útil, empleando estándares como RDF (Resource Description Framework) o SPARQL (SPARQL Protocol and RDF Query Language).
4. Incluir enlaces a otras URI, para poder descubrir otras cosas.

2.5.1.2 *Clasificación por estrellas Linked Data*

Berners Lee, en 2010, define cinco niveles de datos abiertos y vinculados. El objetivo del publicador debe ser conseguir el máximo número de estrellas que califiquen los sets de datos expuestos.

★ Poner a disposición los datos en la web en cualquier formato y con una licencia abierta. Por ejemplo el escaneo de una tabla de datos.

★ ★ Publicar los datos de modo estructurado en formatos legibles por máquina. Por ejemplo una tabla de datos en formato Excel.

★ ★ ★ Publicar los datos además en formatos no propietarios. Por ejemplo datos en formato CSV (Comma Separated Values).

★ ★ ★ ★ Publicar los datos además, a través de estándares abiertos de W3C (RDF y SPARQL) para identificarlos, lo que hace posible que los datos sean enlazados. Por ejemplo datos en formato RDF.

★ ★ ★ ★ ★ Publicar los datos con todos los requisitos anteriores y vincularlos con otros conjuntos de datos.

En el contexto de Open Government Data (OGD) se barajó la posibilidad de añadir una estrella más para aquellos sets de datos que estuvieran descritos con metadatos y que esos metadatos descriptivos estuvieran publicados en un catálogo importante. Ciertamente es que esta estrella no

consta oficialmente como un nuevo nivel de excelencia de Linked Data, pero actualmente, en cualquier iniciativa seria de datos abiertos gubernamentales aparece el requisito de la aplicación de metadatos descriptivos que permitan un nivel agregado de interoperabilidad (Berners-Lee, 2009; Bizer et al., 2009).

Los conceptos de Open Data y Linked Data se confunden en ocasiones. Open Data hace referencia a datos publicados y disponibles bajo una licencia abierta, pero no necesariamente han de estar vinculados (aunque es deseable) a otros datasets ni descritos con lenguajes normalizados. Linked Data no supone a priori la utilización de una licencia abierta; los datos se vinculan por Internationalized Resource Identifiers (IRI) y se utilizan estándares de descripción como Resource Description Framework (RDF). Linked Open Data aúna la vinculación de datos y su exposición mediante licencias abiertas.

En el concepto básico de Linked Data se ha de definir también la noción de Linked Closed Data, que supone que no siempre es posible la publicación en abierto, debido a cuestiones de privacidad, económicas o de cualquier índole. Este subgrupo de datos vinculados tiene unas características propias que se han de considerar, como el grado de restricción del acceso, el establecimiento de sus condiciones económicas (pago automático), los desafíos técnicos a los que obliga la restricción de acceso o la estandarización del acceso en cerrado (Cobden, Black, Gibbins, Carr, & Shadbolt, 2011).

2.5.2 TECNOLOGÍAS LINKED DATA

2.5.2.1 Uniform Resource Identifier URI. Internationalized Resource Identifiers IRI.

En este trabajo se va a utilizar el nuevo concepto de identificador de recursos que ya han establecido como estándar las nuevas recomendaciones del W3C del 2013-2014. En algunos casos se mantiene la expresión URI, pero por una cierta coherencia de los conceptos.

Los identificadores internacionalizados de recursos son expresiones de texto que referencian a recursos de cualquier tipo. El juego de caracteres que utilizan es el UNICODE/ISO 10646 (que permite una inclusión de tipos de texto más amplia que con el juego de caracteres ASCII) y queda definido en detalle en la RFC 3987 (Duerst & Suignard, 2005). Los IRIs han nacido de la mano de la internacionalización que se está llevando a cabo en el seno del W3C y persigue el establecimiento de estándares multilingüísticos que permitan utilizar otros idiomas en los identificadores de recursos. Cierto es que existen algunos problemas de interoperabilidad y por ende de compatibilidad entre URI e IRI, pero esos problemas pueden solventarse mediante la normalización de los IRI según la RFC 3897 (Cyganiak, Wood, & Lanthaler, 2014).

Estos identificadores utilizan el protocolo HTTP, lo que permite establecer el direccionamiento hacia el recurso, identificándolo de modo unívoco, accediendo a la información que contiene y diferenciando los objetos vinculados haciéndolos reconocibles por ordenadores con un lenguaje común independiente de sistemas, aplicaciones u otros entornos tecnológicos. Este sistema de identificación permite la construcción de una red de recursos escalable, lo que favorece la interacción de recursos y ofrece las ventajas de la recombinación de datos e información propias de Linked Data (Bizer et al., 2009; Heath & Bizer, 2011; Hyvönen, 2012a).

La viabilidad del propio sistema descansa en garantizar que los identificadores sean persistentes, manteniendo la vinculación incluso ante el posible movimiento o desaparición de los datos, gestionando la redirecciones posibles y las respuestas de estado de HTTP del servidor (Dirección General de Modernización Administrativa, Procedimientos e Impulso de la Administración Electrónica, 2013).

Existen dos tipos diferentes de IRIs: por un lado aquellos que identifican cosas reales o conceptos, como puede ser una localización, un objeto, un hecho histórico, etc.; por otro lado están los que hacen referencia a documentos o recursos web. Los recursos pueden estar diseñados para ser legibles por máquinas (mediante la descripción en RDF) o por seres humanos (mediante un documento HTML por ejemplo). Los recursos, ya sean conceptos, individuos o propiedades estarán identificados por su propio IRI. De estas consideraciones podemos extraer la consecuencia de que cada recurso ha de ser identificado por tres IRIs distintos: el que vincula al concepto en abstracto, el que vincula al objeto representado por un documento legible por personas y el que vincula a la propiedad o descripción semántica del objeto mediante RDF (Dirección General de Modernización Administrativa, Procedimientos e Impulso de la Administración Electrónica, 2013; Hyvönen, 2012a).

Los recursos identificados por IRIs se denominan referentes. Dado que el IRI tiene una naturaleza universal, cabe la posibilidad de que existan dos IRI que indiquen el mismo referente, el problema que se manifiesta se denomina colisión de IRIs. Por convención es el propietario quien configura el IRI para identificar unívocamente el referente, pero ni aplicaciones ni clientes tienen la necesidad de tener esto en cuenta, lo que puede provocar un problema grave de interoperabilidad. Por tanto identificar el referente es fundamental, siendo deseable la construcción de una ontología global de IRI, pero no parece que esto sea posible a medio plazo. (Cyganiak et al., 2014).

Una cuestión fundamental es que los IRI bajo HTTP, no sólo identifican recursos, sino que deben también obtener información del mismo. De este hecho nace el concepto de IRI desreferenciable, es decir que los clientes HTTP utilizan este protocolo para localizar IRIs y obtener una descripción del recurso cuyo IRI ha sido identificado. La petición de información puede estar referida a un concepto del mundo real, que es requerido por un software de extracción semántica de información, o un documento de carácter informativo legible por humanos y que está presentado en HTML (Auer, Lehmann, Ngonga Ngomo, & Zaveri, 2013; Heath & Bizer, 2011; Hyvönen, 2012a).

Los métodos para desreferenciar IRIs son dos:

1. Método *303 See Other*: el servidor responde a la petición del cliente con un código de respuesta HTTP 303 *See Other*, lo que supone que el recurso no es un documento web, sino un concepto del mundo real y redirige a otro IRI que no es sustituto de la anterior es decir, desde el IRI del concepto nos redirige por ejemplo a un documento web que lo describe (Cyganiak & Sauermann, 2008).
2. Método *hash IRI*: en primer lugar los *hash IRIs* tienen un fragmento especial separado de el IRI mediante el carácter “#”. El cliente quiere recuperar un recurso identificado con un *has IRI*, en ese momento el fragmento es desligado del IRI y posteriormente mediante negociación de contenido se ofrece una representación legible por humanos o una semántica mediante RDF (Cyganiak & Sauermann, 2008).

Por otro lado cabe también identificar ciertas recomendaciones a nivel internacional en cuanto a la elaboración de IRIs:

1. Nombrar los datos mediante IRI es la primera recomendación, aunque obvia es el primer paso, intentando evitar literales en lo posible.
2. No utilizar dominios que no se controlan, pues no se puede modificar la capacidad de respuesta, ni llevar a cabo ninguna política de gestión de enlaces, ni garantizar la persistencia.
3. No se deben crear IRIs con contenidos o infraestructuras que puede ser necesario cambiar en el futuro.
4. Utilizar conceptos naturales (se puede entender como concepto natural, aquellas categorías fácilmente entendibles por el hombre), incluyendo cadenas de texto descriptivo del objeto. Es preciso evitar por ejemplo numeración de serie.
5. Emplear IRIs neutrales, que no muestren las características técnicas de la infraestructura que soporta a los datos, por ejemplo que no muestren extensiones de servidor.
6. Los identificadores de fragmentos son aquellos que siguen a la almohadilla (“#”) y no son fácilmente procesados por los navegadores y por tanto, caben errores de remisión al servidor, siendo únicamente procesados localmente. Por ello deben ser usados con moderación (Cyganiak & Sauermann, 2008; Heath & Bizer, 2011; Wood, Zaidman, Ruth, & Hausenblas, 2014).

En el ámbito de la Unión Europea se ha preparado un informe en el contexto del programa ISA (Interoperability Solutions for European Public Administrations) que pretende ofrecer unos patrones de diseño y de perdurabilidad en la construcción de IRIs que, aunque no tienen valor normativo, si pretenden homogeneizar su diseño con el objetivo de mejorar la interoperabilidad y el intercambio de información a nivel de la EU. Estas mejores prácticas han sido extraídas previo análisis de los casos de uso más exitosos en este ámbito (Archer, Goedetier, & Loutas, 2012). Podemos resumirlos en estos puntos fundamentales:

1. El patrón de uso recomendado es el siguiente:

http://{dominio}/{tipo}/{concepto}/{referencia}

Por dominio se entiende tanto el sector desde el que se construye el IRI, y/o el host, cuando es relevante o una combinación de ambos. Por tipo, el modelo de recurso identificado con un código semántico: doc (documentos), set (datasets), def (conceptos). El concepto hace alusión al conjunto al que hace referencia el IRI, como puede ser una colección de objetos o recursos, la denominación de un “esquema”, etc. La referencia posteriormente especifica estos conceptos.

2. Se recomienda no mencionar a la organización en el IRI, esto es positivo en aquellos casos donde el proyecto u organismo puedan no continuar existiendo.
3. No reflejar los cambios de versión frecuentes en recursos como “esquemas”, ontologías, etc.
4. Si los recursos tienen identificación unívoca, esta debe ser incluida en el IRI.
5. Crear nuevos IRIs para grandes conjuntos de modo automático y unívoco. Se puede establecer un proceso secuencial por identificadores correlativos siempre que los procesos que originan la creación del IRI no se repitan, o si se repiten sean para la expresión de los mismos datos.
6. Evitar las cadenas de consulta en los IRIs del estilo *“?parametro=valor”*
7. Evitar incluir en el IRI extensiones de archivo.
8. Ya se ha hablado de la posibilidad de que un mismo contenido esté reflejado en diferentes tipos de recursos, por ejemplo un documento HTML o uno RDF. En estos casos es mejor la asignación de un IRI diferenciado por ejemplo por la extensión del archivo. Esto no contradice la regla anterior, en tanto en cuanto no dejen de ser IRIs persistentes.
9. Se debe procurar que las diferentes representaciones de un mismo recurso estén vinculadas o relacionadas entre sí.
10. Se recomienda, como se dijo anteriormente, que cuando un IRI identifica un objeto del mundo real, el cual evidentemente no puede ser accedido físicamente, se utilice el código de respuesta HTTP 303 para indicar la ruta del recurso que lo describe.
11. Se recomienda que los servicios de provisión de IRIs sean independientes del creador de los datos de los datos.

2.5.2.2 Modelo de datos de Linked Data: RDF

Frente a iniciativas de vinculación semántica de datos de los grandes gestores de la información como Google, Facebook, Twitter o Amazon, Linked Data nos ofrece una API estándar para la publicación de los datos de modo semántico, esta API pretende ser una plataforma común, un patrón que unifique los proyectos de publicación semántica en la Web (Isaac & Summers, 2005; Wood et al., 2014). Vamos a describir de modo muy breve los diferentes aspectos del modelado de datos semántico. En primer lugar nos centraremos en el modelo gráfico y los vocabularios de representación, posteriormente explicaremos la agrupación natural de los datos y los diferentes modos o formatos de expresión, para finalmente abordar las áreas de discusión y mejora del modelo.

2.5.2.2.1 Modelo gráfico

Las declaraciones RDF describen de modo semántico los objetos y establecen relaciones entre ellos (se entiende en este trabajo que al mencionar RDF se hace genéricamente, haciendo referencia al conjunto de estándares que se incluyen en su infraestructura). Desde un punto de vista técnico estamos ante la típica expresión de entidad-relación de las bases de datos, es decir: entidad/atributo/valor. La comunidad Linked Data identifica estos mismos elementos como sujeto/predicado/objeto, las denominadas tripletas RDF.

Supongamos un caso en el que expresamos gráficamente un elemento de un tesoro. En lenguaje natural diríamos que el concepto “programación” tiene como etiqueta preferida al término “Programación”, y que su tipo es un concepto (*Concept*).

Al describir con RDF esta estructura, proponemos que:

| Sujeto | Predicado | Objeto |
|--------------|--------------------|--------------|
| Programación | Etiqueta preferida | Programación |
| | Tipo | Concept |

Tabla 1 Modelo tabular de tripleta RDF. Fuente: elaboración propia.

Para plasmar la información anterior de un modo más comprensible podemos “dibujar” los elementos en un grafo que relacionará unos elementos con otros. En grafo RDF el sujeto, el predicado y el objeto han sido sustituidos por IRIs que identifican mejor el recurso, o por literales (partes de la descripción que son expresadas mediante el lenguaje natural).

Los grafos o tripletas resultantes nos indican lo mismo que el lenguaje natural o la clasificación en registros de base de datos. Veamos un ejemplo:

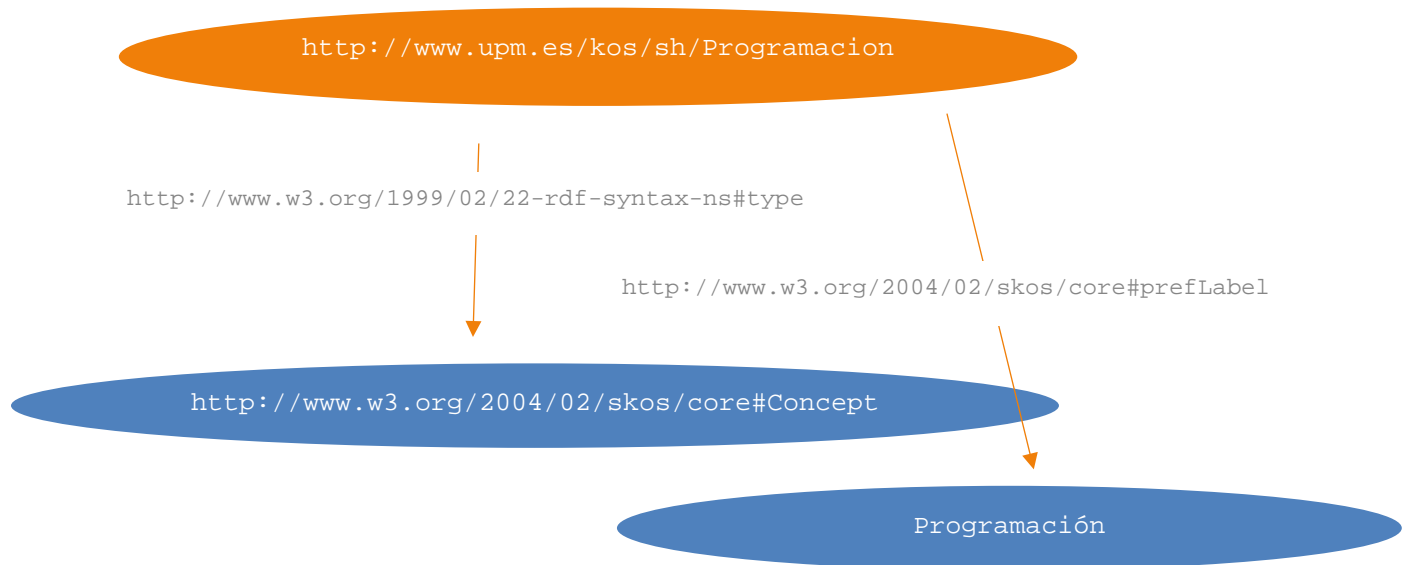


Figura 5 Representación gráfica de una tripleta RDF. Fuente: elaboración propia.

El primer nodo es un IRI que hace referencia al concepto abstracto (sujeto) y describe el concepto del tesoro “Programación” y tiene dos vínculos establecidos mediante IRIs (que representan predicados) que definen la etiqueta preferida del concepto “Programación” (objeto literal) y el tipo de elemento que es *Concept* (objeto), ambos representados por sendos nodos (Gandon & Schreiber, 2014; Isaac & Summers, 2005; Wood et al., 2014).

En el modelo de datos de RDF es importante definir el concepto de nodo en blanco, o nodo sin contenido y que suele utilizarse para enlazar colecciones de elementos (nodos), su función es de distribuidor desde un elemento común. Como se dijo anteriormente, los literales son cadenas de texto que suelen expresar el valor de una propiedad, nombres de personas, localizaciones etc. (Cyganiak et al., 2014).

Los tipos de datos o *datatype* son literales de RDF y representan valores de datos determinados. Estos tipos de datos pueden consistir en un espacio léxico (una cadena de texto), un espacio de valor o un mapeo léxico valor. Los tipos de datos XML Schema, son compatibles con RDF: *xsd:string*, *xsd:boolean*, *xsd:decimal* o *xsd:integer*, y se expresan como:

http://www.w3.org/1999/02/22-rdf-syntax-ns#string

`http://www.w3.org/1999/02/22-rdf-syntax-ns#boolean`

`http://www.w3.org/1999/02/22-rdf-syntax-ns#decimal`

`http://www.w3.org/1999/02/22-rdf-syntax-ns#integer`

2.5.2.2.2 Vocabularios semánticos

Una parte fundamental del modelo de datos RDF son los vocabularios. RDF aumenta su capacidad de estructuración de la información si se combina con vocabularios específicos para la descripción semántica. Estos vocabularios RDF son definiciones de términos utilizados para efectuar los vínculos entre los diversos elementos de una descripción RDF (Wood et al., 2014). RDF Schema (RDFS), por ejemplo, es un lenguaje que ayuda a expresar y jerarquizar semánticamente los datos. Los principales elementos de RDFS son:

1. Las Clases: para describir jerarquía de elementos.
2. Las Propiedades: describen características o relaciones de los recursos.
3. Limitaciones o restricciones a las propiedades: indican qué propiedades y atributos pueden tener un grupo de recursos (Brickley & Guha, 2014a; Schreibe & Raimond, 2014).

Los vocabularios RDF más utilizados son estándares que se han formado por la amplia utilización y desarrollo de la comunidad y son el pilar básico de la reutilización de la información. Estos vocabularios utilizan IRI bajo HTTP lo que permite su utilización en red, y son identificados como colecciones de términos o espacios de nombres.

Podemos enumerar algunos ejemplos de los denominados vocabularios autorizados, aquellos que presentan una utilización más extendida y reconocida:

Tabla 2 Principales vocabularios semánticos. Fuente: elaboración propia.

| Vocabulario | Prefijo | IRI |
|-------------------------|--------------|--|
| Dublin Core Terms | dct: | <code>http://purl.org/dc/terms/</code> |
| RDF | rdf: | <code>http://www.w3.org/1999/02/22-rdf-syntax-ns#</code> |
| Dublin Core Elements | dc: | <code>http://purl.org/dc/elements/1.1/</code> |
| Data Catalog Vocabulary | dcat: | <code>http://www.w3.org/ns/dcat#</code> |
| Void | void: | <code>http://rdfs.org/ns/void#</code> |
| SKOS | skos: | <code>http://www.w3.org/2004/02/skos/core#</code> |

El prefijo de un espacio de nombres es una especie de intermediario de los IRIs a las que representa y sustituye. Permiten una mejor visibilidad de la descripción y utilizan una sintaxis tipo atributo que declara la asociación del prefijo con el espacio de nombres de esa referencia IRI.

2.5.2.2.3 Conjuntos de datos

Ya hemos dicho anteriormente que los triples RDF se expresan mediante grafos. Estos grafos que representan declaraciones RDF suelen estar agrupados en colecciones, lo que RDF Primer 1.1 (2014) denomina *multiple graphs* o *RDF datasets*. Estos datasets están identificados mediante un IRI o nodo en blanco. La recomendación RDF denomina *named graphs* a esos conjuntos de declaraciones RDF y *graph name* a los IRIs o nodos en blanco. Al grafo que no está identificado se le denomina *default graph*. Esta construcción de RDF 1.1 facilita la descripción de grupos de datos representados semánticamente, ofreciendo un modelo acorde a las necesidades de identificación de los conjuntos de datos, vehículo fundamental para la asignación de datos de procedencia, o de copyright. Hay que decir que dentro de la semántica del conjunto de datos, el *named* no es imprescindible para denominar al grafo, más bien se empareja sintácticamente con él (Schreibe & Raimond, 2014).

2.5.2.2.4 Serialización

La representación de los datos mediante grafos RDF no es el sistema adecuado para la expresión de declaraciones cuando estamos ante amplios conjuntos de datos y con una configuración compleja de sus relaciones. La serialización permite la transformación del lenguaje gráfico a lenguaje máquina, utilizando lenguajes codificados. RDF tiene varias formas diferentes de representación, cada una de ellas con sus ventajas e inconvenientes y diseñadas para un marco de representación determinado. Así RDF/XML es la sintaxis más habitual en el ámbito empresarial, mientras que Turtle es uno de los formatos más fáciles de leer por las personas.

RDF 1.1 (2014) ya no establece RDF/XML como único formato recomendable, los nuevos formatos más legibles por humanos son considerados parte del modelo de datos de RDF y por ello plenamente admisibles para su utilización (Hyvönen, 2012b; Schreibe & Raimond, 2014; Wood et al., 2014).

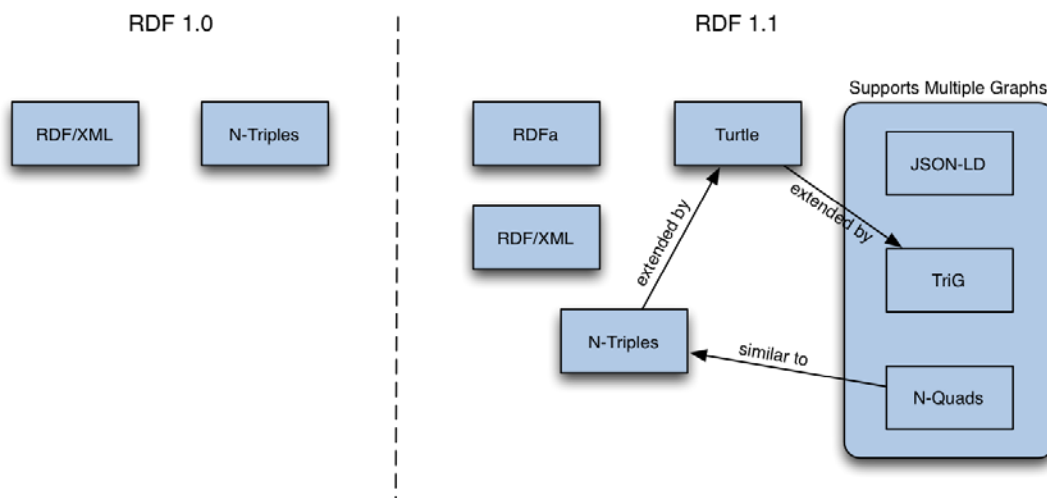


Figura 6 Nuevas posibilidades de serialización RDF. Fuente: Recomendación W3C RDF 1.1 2014

Como se puede observar en la figura 6, los formatos de serialización están evolucionando para adaptarse a los cambios del modelo. La denominada “familia Turtle” se compone del básico N-Triples, N-Quads (su evolución para adaptarse a grafos múltiples), TriG que evoluciona directamente de Turtle para dar soporte a los grafos múltiples y el propio formato Turtle.

2.5.2.2.4.1 N-Triples

Presenta cada triple en una línea donde se separan el sujeto, el predicado y el objeto mediante un espacio en blanco. El final del triple termina en punto. Los IRIs están encerrados entre marcas “< >”. El tipo de datos para literales se introduce con ^^ como delimitador y el atributo XML de tipo de dato. (Beckett, 2014). El ejemplo siguiente se refiere al grafo incluido en la figura 5.

```

<http://www.upm.es/biblioteca/kos/sh/programacion>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://www.w3.org/2004/02/skos/core#Concept> .

```

```

<http://www.upm.es/biblioteca/kos/sh/programacion>
<http://www.w3.org/2004/02/skos/core#prefLabel> "Programación" .

```

2.5.2.2.4.2 N-Quads

Sintaxis similar a N-Triples, pero con posibilidad de serializar *datasets*, añadiendo al final de la línea el *default graph* que especifica al dataset. La secuencia de serialización en N-Quads comprende el sujeto, el predicado, el objeto y un grafo que etiqueta la tripleta RDF e indica la pertenencia a un dataset.

Las diferentes partes de la línea de serialización N-Quad pueden ser IRIs, nodos en blanco o literales (los IRIs encerrados en marcas "<>", los nodos en blanco precedidos de " _: " y los literales entre comillas) separados por un espacio en blanco y finalizado por un punto (Carothers, 2014). Siguiendo con el ejemplo anterior:

```
<http://www.upm.es/biblioteca/kos/sh/programacion>  
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
<http://www.w3.org/2004/02/skos/core#Concept>  
  
<http://www.upm.es/biblioteca/sh/UPMSubjectHeadings.ttl> .  
  
<http://www.upm.es/biblioteca/kos/sh/programacion>  
<http://www.w3.org/2004/02/skos/core#prefLabel> "Programación"  
  
<http://www.upm.es/biblioteca/sh/UPMSubjectHeadings.ttl> .
```

2.5.2.2.4.3 Turtle

Es el formato de serialización diseñado para una lectura más sencilla por humanos. A día de hoy es el formato de serialización RDF más utilizado por la comunidad de la Web Semántica y los desarrolladores de datos bajo Linked Data. Algunas pautas principales de su sintaxis son las siguientes (Beckett, Berners-Lee, Prud'hommeaux, & Carothers, 2014; Wood et al., 2014):

1. Sujeto, predicado y objeto van separados por espacio en blanco, terminando la tripleta con un punto.
2. Los comentarios se escriben tras el signo "#".

3. Los sujetos pueden tener varios predicados con sus objetos correspondientes. Para abreviar esa expresión el sujeto sólo se describe una vez, seguido de varios conjuntos predicado-objeto separados por el signo “;”. Tanto se puede decir de los objetos, en cuyo caso van separados por el signo “,”.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sh: <http://www.upm.es/biblioteca/kos/sh/> .

<http://www.upm.es/biblioteca/kos/sh/programacion>
  a skos:Concept ;
  skos:prefLabel "Programación" ;
  skos:altLabel "Simbolización" , "Categorización" ,
  "Codificación" .
```

4. Los IRIs en Turtle se presentan entre “< >”. Existen dos tipos, absolutas o relativas: las primeras describen la URI completa, las segundas se representan con respecto a una URI base introducida con “@ base”.

```
@base <http://www.upm.es/biblioteca/kos/sh/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<#programación>
  a skos:Concept ;
<#sistemas>
  a skos:Concept .
```

5. Por una cuestión de economía de representación el token “a” dentro de la declaración Turtle representa a la IRI (propiedad) *rdf:type* o lo que es lo mismo *http://www.w3.org/1999/02/22-rdf-syntax-ns#type* (es posible la utilización también de la forma *rdf:type*).

```
@base <http://www.upm.es/biblioteca/kos/sh/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<#programación>
  a skos:Concept ;
<#sistemas>
  a skos:Concept .
```

6. La etiqueta “@prefix” indica el identificador del *namespace* seguido del signo “:”. Esto permite declarar la utilización del *namespace* y la utilización a discreción del prefijo. Semánticamente sustituye al IRI entero más el valor local.

```
@base <http://www.upm.es/biblioteca/kos/sh/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<#programación>

a skos:Concept . #representación con prefijos

<#sistemas>

http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/2004/02/skos/core#Concept .
#la representación lineal con IRI tiene el mismo valor
descriptivo
```

7. Los literales utilizan diferentes signos para identificar su tipología, los principales son: “” para literales textuales, “^^” para tipos de datos, “@” para etiquetas.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sh: <http://www.upm.es/biblioteca/kos/sh/> .

<http://www.upm.es/biblioteca/kos/sh/programacion>
  a skos:Concept ;
  skos:prefLabel "Programación"^^xsd:string ; #tipo de dato
  skos:altLabel "Simbolización", "Codification@en" ,
               "Codificación@es" .
```

8. Los nodos en blanco se identifican con el conjunto de signos “_:”. Partimos del siguiente grafo:

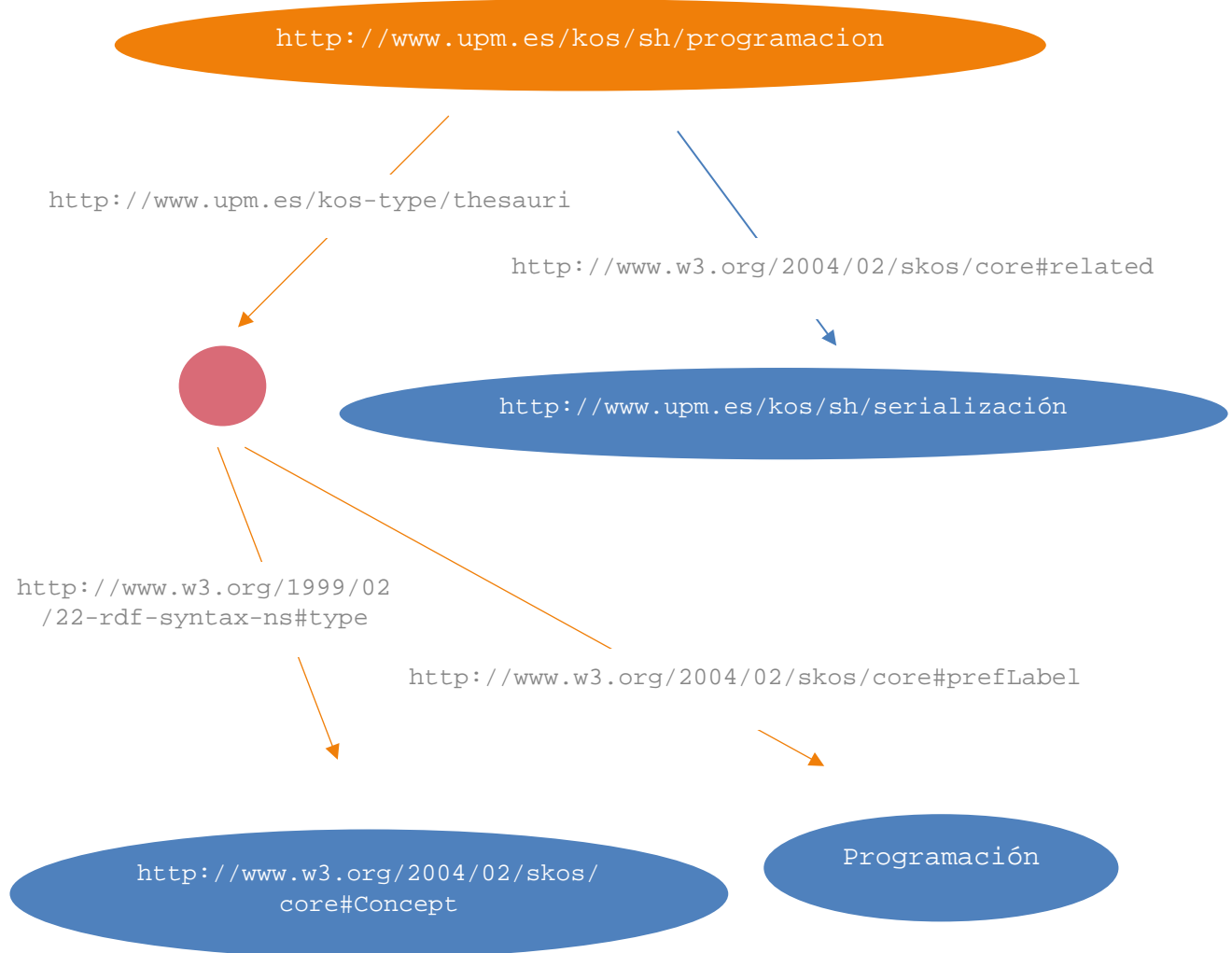


Figura 7 Representación mediante grafos de una declaración RDF con nodo en blanco. Fuente: elaboración propia.

Su representación en la sintaxis Turtle se serializa del siguiente modo:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sh: <http://www.upm.es/biblioteca/kos/sh/> .
@prefix kostype: <http://www.upm.es/biblioteca/kos_type/>

<http://www.upm.es/biblioteca/kos/sh/programacion> .
skos:related sh:serialización
_:bnode1 a <http://www.w3.org/2004/02/skos/core#Concept> .
_:bnode1 skos:prefLabel "Programación" .

kostype:thesauri _:bnode1 .
```

9. En los nodos en blanco pueden anidarse los elementos mediante el símbolo "[]"

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sh: <http://www.upm.es/biblioteca/kos/sh/> .
@prefix kostype: <http://www.upm.es/biblioteca/kos_type/>

<http://www.upm.es/biblioteca/kos/sh/programacion> .
skos:related sh:serialización ;
kostype:taxonomy [
a skos:Concept ;
skos:prefLabel "Programación"
] .
```

2.5.2.2.4.4 TriG

Es una extensión del lenguaje Turtle, cuyo diseño responde a la necesidad de descripción de datasets en un contexto de sencillez, utilizando un lenguaje compacto y natural. Los triples se describen entre corchetes { } y son etiquetados por un grafo por defecto que puede estar representado por un IRI o un nodo en blanco. Un único dataset (grafos múltiples) puede estar constituido por un único grafo.

Las denominaciones de estos grafos no son repetibles en el dataset (Bizer & Cyganiak, 2014). Veamos un ejemplo con un grafo por defecto y dos grafos nominados:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sh: <http://www.upm.es/biblioteca/kos/sh/>
```

```
#grafo por defecto
{
  <http://www.upm.es/biblioteca/kos/sh/Programacion> a
  skos:Concept .
  <http://www-upm.es/kos/sh/Sistemas> a skos:Concept .
}
```

```
<http://www.upm.es/biblioteca/kos/sh/Programación>
{
  _:bnode1 skos:prefLabel "Programación" ;
  _:bnode1 skos:altLabel "Simbolización";
  _:bnode1 skos:related "Codificación" .
}
```

```
<http://www.upm.es/biblioteca/kos/sh/sistemas>
{
  _:bnode2 skos:prefLabel "Sistemas" ;
  _:bnode2 skos:altLabel "System@en" ;
  _:bnode2 skos:related "Estructuras" .
}
```

2.5.2.2.4.5 RDFa

Es un formato de serialización que permite la inclusión de RDF en HTML mediante atributos. Los datos se estructuran mediante códigos dentro del formato de marcado y los atributos HTML suministran la suficiente información sobre esos datos para que los parsers los procesen y generen tripletas RDF. RDFa por tanto permite la legibilidad por máquina de los datos.

Para una mejor gestión del marcado, RDFa tiene también un sistema de acortamiento de IRIs denominado *Compact URI Expressions (CURIE)*. Este sistema es un subconjunto de los *QNames* de RDF/XML permitiendo una serie de recursos descriptivos que mejoran la recuperabilidad de los contenidos, y optimizan los sistemas de búsqueda SEO. La utilización de IRIs frente a palabras en RDFa aporta ventajas a los contenidos: los hace más portables, interoperables y suma coherencia al sistema, además los IRI restan ambigüedad terminológica, sobre todo a nivel de comprensión máquina (Herman, Adida, Sporny, & Birbeck, 2013).

Además, RDFa incluye una técnica para la descripción semántica de URI mediante un etiquetado que permite su legibilidad por máquina. Veamos un ejemplo:

```
<p>
El vocabulario presentado se describe bajo las recomendaciones de la
<a property="http://www.w3.org/TR/swbp-skos-core-spec
  href=http://www.w3.org/TR/skos-reference/"> W3C
</a> referente al sistema SKOS para la descripción semántica de
vocabularios controlados
</p>
```

RDFa permite la declaración de vocabularios determinados mediante la etiqueta *vocab*. En la descripción siguiente se indica que se van a utilizar las etiquetas del vocabulario SKOS para la descripción semántica del contenido de una web determinada.

```
<body vocab="http://www.w3.org/TR/skos-primer/">
```

Las etiquetas más habituales en el formato RDFa son las siguientes:

1. *Resource* para poder indicar distintas entradas descriptivas cada una con su correspondiente descripción.
2. *Typeof* sirve para introducir nuevos datos no declarados todavía. Para utilizarlos necesitamos introducir los vocabularios correspondientes a esos términos.
3. *Property* para establecer la relación entre cualquier sujeto que se pretenda describir y los recursos que pueden considerarse objetos de una tripleta RDF.
4. *Vocab* tal y como se dice arriba, permite la introducción de ciertos términos a través de la identificación del vocabulario que se precise.
5. *Prefix* se utiliza de modo análogo a RDF, permitiendo tras la identificación del *namespace* acortar las IRI cuando sea necesario.

La técnica de incluir descripciones en textos marcados con HTML no es nueva, tanto los microdatos, como los microformatos ya son tradicionalmente utilizados para definir la estructura semántica interna de los textos electrónicos. Hoy por hoy, la especificación más importante junto a RDFa para estos cometidos es *Schema.org*, iniciativa de descripción semántica de los gigantes de la búsqueda de información como Google, Yahoo o Bing. La diferencia fundamental entre RDFa y *Schema.org* es que la primera utiliza IRI y se integra de modo natural en Linked Data, mientras que *Schema.org* se basa en la utilización de esquemas específicos utilizados como etiquetas HTML. (Herman et al., 2013; Hyvönen, 2012b).

2.5.2.2.4.6 JSON-LD

Entre los desarrolladores Web se utiliza un lenguaje de programación desarrollado en Java que incluye soporte para Linked Data. Se trata de JSON-LD, una evolución de JSON, que admite la descripción semántica de los datos y que permite una serialización RDF interna. Este formato también admite la descripción de grafos múltiples o datasets. Su sintaxis es combinable de modo natural con los sistemas distribuidos que utilizan JSON, permitiendo utilizar Linked Data en entornos de programación web, en sistemas de almacenamiento JSON y la generación de nuevos servicios de interoperabilidad Web. Podemos describir de un modo simple la sintaxis de JSON-LD (Sporny, Longley, Kellogg, Lanthaler, & Lindström, 2014; Wood et al., 2014):

1. La descripción se encabeza con el objeto de contexto “@context”. Aquí se indican los *namespaces* que fijan los prefijos, lo que permite utilizar IRIs cortas.
2. En el caso de datasets, la agrupación de nodos puede hacerse mediante la utilización de “@graph”
3. Con “@id” se identifican los objetos descritos con un IRI o un nodo en blanco. Son en realidad sujetos RDF.
4. Los predicados vienen reflejados mediante *@type*, tipo especial *rdf:type*, o mediante simples cadenas de texto entrecomilladas o con la expresión *rdfs:label*
5. Las expresiones generales se hacen mediante los símbolos “{ }” y para anidaciones de triples se utiliza “[]”.

Utilizando los ejemplos anteriores, podemos presentar un modelo de descripción de JSON-LD

```
{
  "@context": {
    "rdf" : http://www.w3.org/1999/02/22-rdf-syntax-ns# ,
    "skos" : http://www.w3.org/2004/02/skos/core# ,
    "sh" : http://www.upm.es/biblioteca/kos/sh/
  },

  "@graph":
  [
```

```
{
  "@id": "http://www.upm.es/biblioteca/kos/sh/programacion" ,
  "@type": "http://www.w3.org/2004/02/skos/core#Concept" ,
},
{
  "@id": "http://www.upm.es/biblioteca/kos/sh/sistemas" ,
  "@type": "http://www.w3.org/2004/02/skos/core#Concept"
}
]
}
```

2.5.2.2.4.7 RDF/XML

Este formato de serialización siempre ha tenido detractores debido a su complejidad. También ha perdido su papel protagonista como formato de serialización normativo, pues hoy ya se considera plenamente recomendable las descripciones en formatos como Turtle. En cambio RDF/XML sigue siendo el formato que más se adecua a la estructura empresarial de la información, pues sus infraestructuras y sistemas siguen habitualmente tecnologías basadas en XML.

La transformación de los grafos RDF se efectúa mediante los términos XML: nombres de elementos y sus contenidos, y nombres de atributos y sus valores. RDF/XML utiliza QNames para representar IRIs, por un lado es habitual que el QName esté representado por un término corto que equivale al *namespace* y por otro está la parte local del QName que toma el valor que se asigne según la descripción. Estos QNames sirven para acortar los IRI en los nodos en los que se utiliza, pudiendo ser sujetos, objetos y valores de un atributo. Los literales sólo pueden ser objetos, pudiendo ser ya contenido de texto como elemento XML o valores de un atributo.

Se exponen a continuación varias pautas básicas del formato de serialización RDF/XML según la última recomendación publicada, febrero 2014, aunque hay un borrador que redefine algunos aspectos y que a fecha de terminación de este trabajo no ha sido publicado (Gandon & Schreiber, 2014).

1. El comienzo de una serialización en RDF/XML puede contener la declaración XML (`<?xml version="1.0"?>`), indicando la versión y la codificación del contenido.
2. La cabeza del documento continua con el elemento XML `rdf:RDF`, que convencionalmente se utiliza para incluir los *namespaces*, aunque su utilización no es obligatoria.

3. El siguiente elemento es *rdf:Description*, que introduce el recurso a describir y engloba una o más sentencias sobre ese recurso. La identificación se hace con un atributo *rdf:about*.
4. El elemento *rdf:type* puede ser utilizado para introducir los predicados de las sentencias. Los predicados pueden ser introducidos por *QNames* e IRIs.
5. Los objetos pueden ser descritos mediante *rdf:resource* si se precisa incluir el recurso con su referencia IRI completa.
6. Los objetos pueden ser presentados como en el resto de formatos, como literales, presentándose como contenido definidor de los predicados y no como atributos de los mismos. Los literales se declaran con el elemento *rdf:parseType="Literal"*.
7. Las secuencias de datos en orden requieren del elemento *rdf:Seq*, cuando el orden no es prescrito podemos utilizar el elemento *rdf:Bag* y si necesitamos incluir una secuencia cuyos datos son alternativos utilizaremos *rdf:Alt*. Para la identificación de la lista de recursos se utilizan los elementos *rdf:li* o *rdf:_n* (donde “n” es una secuencia numérica correlativa).
8. El elemento *rdf:datatype* permite la utilización de literales especiales que constan de una parte textual y un IRI.
9. El elemento *rdf:nodeID* permite, mediante la asignación de un identificador al nodo en blanco, su reutilización en otras partes de la descripción.
10. La conjunción de los elementos *rdf:ID* y *xml:base* permite establecer un sistema de acortamiento de IRI. Esto se produce mediante la utilización del atributo *xml:base* que identifica la IRI que luego será representada mediante *rdf:id* y el atributo local. Este atributo local es un término del ámbito de la IRI de *xml:base*, y sólo puede definirse una vez.
11. El elemento *rdf:parseType="Collection"* es un atributo de un predicado que permite describir múltiples elementos los cuales están cualificados, es decir se definen como un grupo de recursos limitado y caracterizado por un contenido semánticamente común.
12. Con el elemento *rdf:parseType="Resource"* se puede omitir *rdf:Description*.
13. El elemento *rdf:ID* puede utilizarse para reificar una declaración completa RDF en otra que la contiene. Para ello se construye un identificador *rdf:ID* y un término que identifica el triple embebido en el dominio de un IRI (base) declarado con *xml:base*

Representación de RDF/XML del grafo de la figura 6

```
<?xml version="1.0"?>
  <rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:kostype="http://www.upm.es/biblioteca/kos_type/"

    <rdf:Description
      rdf:about="http://www.upm.es/biblioteca/kos/sh/programacion"
        skos:related
      rdf:resource="http://www.upm.es/biblioteca/kos/sh/serializacion"/
    >
      <kostype:thesauri>
        <rdf:Description
          rdf:type
            rdf:resource="http://www.w3.org/2004/02/skos/core#Concept">
          <skos:prefLabel="Programación" />
        </rdf:Description>
      </kostype:thesauri>
    </rdf:Description>
  </rdf:RDF>
```

2.5.2.2.5 RDFS

RDF permite la utilización de dos lenguajes de ontologías para un mejor modelado de la información: RDFS (RDF Schema) y OWL (*Web Ontology Language*).

RDFS ofrece un vocabulario de modelado de datos que mejora de la expresividad de los recursos descritos. Permite la expresión semántica de grupos de elementos conjuntados y establece un sistema de relaciones más completo. Sus componentes se utilizan para la definición de las características como dominios y rangos de propiedades

RDFS describe recursos clasificándolos en clases y los elementos que las integran, las instancias de clase, permitiendo la organización jerárquica de dichos elementos. La clase se distinguen del grupo de instancias que la integran y que en conjunto se denominan “extensiones de clase”. Las clases en RDFS se define con *rdfs:Class*. El resto de elementos de clase RDFS son: *rdfs:resource*, la clase recursos, *rdfs:Literal*, para la clase literales, *rdfs:Datatype*, para la clase de tipo de datos, *rdf:langString*, para la clase de las etiquetas textuales de idioma, *rdf:HTML*, para la clase de valores literales de HTML, *rdf:XMLLiteral*, para la clase de valores literales de XML y *rdf:Property*, para la clase de propiedades (Brickley & Guha, 2014a).

Las propiedades RDF (*rdf:Property*) establecen relaciones entre el sujeto y el objeto de la tripleta y pueden organizarse jerárquicamente (*rdfs:subPropertyOf*). Los elementos de propiedad en RDFS son: *rdfs:range*, instancia de *rdf:Property* que se utiliza para indicar que los valores de una propiedad son instancias de una o más clases; *rdfs:domain*, que indica que cualquier recurso que

tiene una propiedad es instancia de una o más clases; *rdf:type*, que indica que un recurso es una instancia de una clase; *rdfs:subClassOf*, que indica que una clase es una subclase de otra; *rdfs:subPropertyOf*, que establece la jerarquía de propiedades; *rdfs:label*, que proporciona una versión legible por humanos del nombre del recurso y *rdfs:comment*, que proporciona una descripción legible por humanos del recurso (Brickley & Guha, 2014b).

2.5.2.2.6 OWL 2

OWL 2 (Lenguaje de Ontologías Web, revisión de OWL 1) permite el procesamiento robusto de la información, interpretándola y representándola de modo más eficaz que los modelos de la familia RDF. OWL 2 está diseñada y adaptada al desarrollo de ontologías compatibles vía web.

En la figura 8 se muestra la disposición de componentes del lenguaje OWL 2 y las relaciones posibles entre ellos. La elipse central representa la noción abstracta de una ontología, los satélites exteriores representan las diversas sintaxis que permiten la serialización y el intercambio de información con la ontología, en la parte de abajo se representan dos especificaciones semánticas definitorias de las ontologías procesadas con OWL 2.

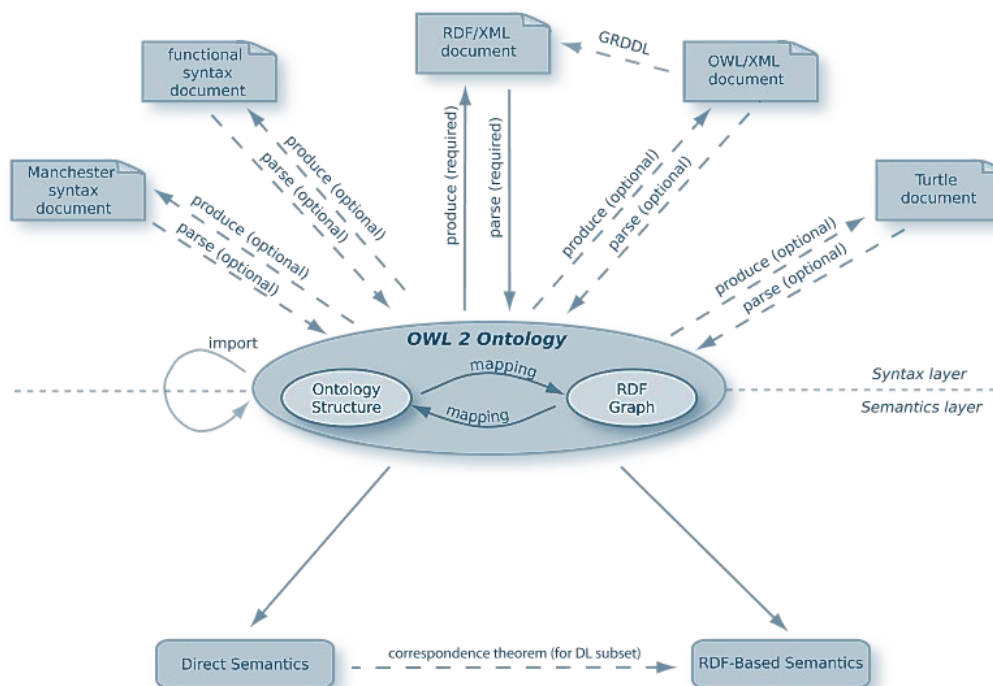


Figura 8 Estructura de OWL 2. Fuente: Recomendación W3C OWL 2, 2012

OWL 2 es capaz de extraer consecuencias de ciertas declaraciones anteriores efectuadas, adoptando el modelo lógico de que una declaración es verdadera si aquellas que la soportan también lo son. Permite que las aplicaciones que analizan su estructura infieran conocimiento, estas aplicaciones son los *reasoners* herramientas para captura automática del conocimiento. Como lenguaje de representación del conocimiento, OWL 2 se estructura sobre tres pilares fundamentales:

1. Los axiomas: enunciados básicos.
2. Las entidades: elementos que hacen referencia a objetos reales.
3. Las expresiones: que son las combinaciones de entidades formando complejas descripciones de axiomas.

En OWL 2 se identifican objetos como individuos, categorías como clases y relaciones como propiedades. Las propiedades en OWL 2 se pueden subdividir:

1. Las propiedades de objetos que relacionan objetos con otros objetos.
2. Las propiedades de tipos de datos que asignan valores a esos datos.
3. Las propiedades de las anotaciones que codifican la información de la propia ontología.

Las entidades están nominadas y pueden ser agrupadas para obtener expresiones más complejas mediante “constructores” (Brickley & Guha, 2014a; Hitzler, Krösch, Parsia, & Rudolph, 2012).

Los principales elementos de la sintaxis OWL2 son los siguientes:

1. Clases e instancias: Las clases representan a grupos de individuos (individuales) cuya unión se establece por pertenecer al campo semántico de un concepto. Las instancias son los elementos componentes de las clases. Su expresión en OWL 2 y Turtle es la siguiente:

```
ClassAssertion( :Libro :El Quijote )  
  
:El Quijote rdf:type :Libro .
```

2. Unas clases pueden englobar semánticamente a otras. Entre ellas se establece una relación de jerarquía expresada mediante el axioma subclase, que tiene la característica de la transitividad, es decir en una jerarquía de tres niveles, la clase tercera es instancia de la segunda y también de la primera clase:

```
SubClassOf( :Libro :Documento )  
  
:Libro rdfs:subClassOf :Documento .
```


3. Si dos clases se refieren al mismo conjunto de individuos, la representación OWL 2 es:

```
EquivalentClasses( :Libro :Obra )  
  
:Libro owl:equivalentClass :Obra .
```

4. Cuando dos individuos pertenecen a clases excluyentes, OWL 2 lo expresa mediante la clase de disjunción. Es conveniente expresar este tipo de clases, pues en muchos casos los razonadores no establecen por si mismos la relación de disjunción:

```
DisjointClasses( :Novela :Ensayo )  
[ ] rdf:type owl:AllDisjointClasses ;  
 owl:members ( :Novela :Ensayo ) .
```

5. OWL 2 permite también describir las relaciones entre individuos y lo hace a través de propiedades. Es importante respetar la dirección de la relación, pues perfectamente puede no ser recíproca:

```
ObjectPropertyAssertion( :escritoPor :El Quijote :Cervantes )  
:El Quijote :escritoPor :Cervantes.
```

6. Cabe también la expresión de jerarquía entre las propiedades, al igual que entre las clases. También es posible el mismo procedimiento en las equivalencias:

```
SubObjectPropertyOf( :tieneAutor :tieneAutor_secundario )  
:tieneAutor_secuncario rdfs:subPropertyOf :tieneAutor .
```

7. OWL 2 permite, como se ha dicho, inferir conocimiento. Si dos individuos están conectados por una propiedad puede extraerse información sobre dominio y rango:

```
ObjectPropertyDomain( :escritoPor :Autor )  
ObjectPropertyRange( :escritoPor :Obra )  
  
:escritoPor rdfs:domain :Autor ;  
 rdfs:range :Obra .
```

8. La expresión semántica puede necesitar que indiquemos que un individuo es igual o diferente a otro. Este conocimiento no lo infiere OWL 2 y es necesario introducirlo.

```
DifferentIndividuals( :Cervantes :Lope_de_Vega )  
:Cervantes owl:differentFrom :Lope_de_Vega .  
  
SameIndividual( :Cervantes :Miguel_de_Cervantes )  
:Cervantes owl:sameAs :Miguel_de_Cervantes.
```

9. Los individuos pueden tener asociados valores que definen y especifican la información que proveen. Estos valores de datos son introducidos por OWL 2 a través de las propiedades de valores de datos, algunos de los cuales pertenecen al esquema XML de tipos de datos (Hitzler et al., 2012).

```
DataPropertyAssertion( :Cumpleaños :Cervantes "09-29-1547"^^xsd:date )
:Cervantes :Cumpleaños 09-29-1547 .
```

2.5.3 MODELO DE SERVICIO DE DATOS EN LINKED DATA

Los datos vinculados están publicados en servidores que deben proveer los recursos que los clientes les solicitan; un complejo mecanismo rige esa negociación de clientes y servidores: la negociación de contenido bajo el protocolo HTTP.

El proceso de comunicación comienza con la petición de recursos por parte de un cliente determinado, que transmite la petición GET mediante un encabezado HTTP *Content-Type*, que recoge un tipo de contenido que se solicita al servidor. La petición puede ser por ejemplo un recurso cuya especificación *Content-Type* es del tipo *application/rdf+xml*, en ese caso se espera del servidor un recurso en RDF que esté formateado con RDF/XML y si es encontrado, el servidor remite el recurso con el código de estado 200 OK *Content-Type application/rdf+xml*. En la tabla siguiente se describen los principales tipos de contenido en relación con su codificación como tipos de datos.

Tabla 3 Principales tipos de datos para solicitudes de contenido. Fuente: elaboración propia.

| Formatos de serialización | Tipo de datos de solicitud |
|---------------------------|----------------------------|
| Turtle: | text/turtle |
| RDF/XML | application/rdf+xml |
| RDFa | text/html |
| JSON-LD file | application/ld+json |
| N-Triples | application/N-Triples |

Los tipos de contenido (tipos MIME) que se emplean en el mecanismo de negociación están definidos bajo registro en el Internet Assigned Numbers Authority (IANA)

2.5.4 CICLO DE VIDA DE LOS DATOS BAJO TECNOLOGÍAS LINKED DATA

Linked Data es un complejo proceso con etapas diversas y no excluyentes, que se retroalimenta y regenera en una constante evolución. El ciclo de vida de los datos bajo tecnologías Linked Data no es un proceso estático, ni completamente estandarizado, por lo que los ítems que lo componen pueden variar de ciertos usos a otros.

La gestión de datos vinculados ofrece varias versiones del ciclo de datos Linked Data. Podemos enumerar algunas de ellas: en primer lugar la iniciativa “lod2” dentro del 7º Programa Marco de Innovación e Investigación de la UE; la propuesta de Villazón Terrazas y otros (2012), del Ontology Engineering Group (OEG) de la Universidad Politécnica de Madrid; la diseñada por Wood y otros (2014); y las recomendaciones de buenas prácticas para la publicación de Linked Data del W3C.

Aquí se expone únicamente un esquema introductorio, pues el desarrollo de las etapas de gestión de datos vinculados se hará con más profundidad durante el desarrollo de este trabajo. Se han escogido las que se consideran mejores prácticas en el conjunto de ellas. (Auer et al., 2013; Bauer & Kaltenböck, 2012; Hyland, Ateamezing, & Villazón-Terrazas, 2014; Hyvönen, 2012b; Villazón-Terrazas, Vilches-Blázquez, Corcho, & Gómez-Perez, 2011; Wood et al., 2014).

1. Adquisición, selección y preparación de los datos: es importante seleccionar conjuntos de datos que sean por su interés o valor reutilizables; lo fundamental es establecer un criterio de selección y clasificación correcto. Cabe la posibilidad de que tengamos que extraer los datos desde sistemas de datos estructurados o desestructurados, desde procesos Big Data a bases de datos relacionales, pasando por datos de actividad de la Administración. En esta parte del proceso, también podemos incluir la generación de IRIs, aunque en algunos procesos esto se hace durante el modelado de datos. También se incluyen en esta fase las tareas de limpieza y clasificación de los datos, además del estudio de las licencias que acompañarán al dataset; y como último punto importante, el estudio del contexto del caso concreto de gestión de datos vinculados: si estamos ante proyectos Open Data u Open Government Data (OGD); o quizás una publicación Linked Open en el ámbito de las bibliotecas o los museos, etc. Otra recomendación importante se refiere a la prospección de las demandas de publicación de datos, en ciertos casos tendremos que recabar e informar a nuestros *Stakeholders* o demandantes de servicios y verificar las necesidades que tienen sobre nuestros datos.

2. Modelado de datos: según las recomendaciones y buenas prácticas del W3C, el proceso de modelado de los datos trata de “capturar” la sustancia de los datos y establecer relaciones semánticamente pertinentes con otros conjuntos de datos. El modelado debe conseguir representar el conocimiento del área definida de estudio, para ello es necesario un lenguaje de modelado específico, un vocabulario estandarizado, o incluso el diseño local de nuestro propio vocabulario u ontología. Dado que las relaciones semánticas se pueden establecer entre datasets, el modelado de los metadatos descriptivos de conjuntos de datos adquiere una especial relevancia y es muy importante para establecer la interoperabilidad global. El proceso de asignación de licencias puede verse modificado en este punto: depende de los condicionantes de publicación que pueden requerir, desde licencias apenas restrictivas, a la publicación de datos no abiertos (Linked Closed Data).
3. Transformación de los datos: los datos han de ser convertidos (serializados) mediante los diferentes lenguajes descriptivos de RDF. Tras el modelado, se efectúa la serialización mediante el lenguaje escogido y de acuerdo al vocabulario que hayamos creado o reutilizado. La serialización automática mediante el mapeo de los datos, permite la conversión de grandes datasets en instancias RDF que mantienen la semántica de los datos y su recuperabilidad en las búsquedas. Existen muchas herramientas preparadas para efectuar esta tarea, R2RML es casi un estándar para el mapeo desde bases de datos relacionales. También la transformación desde *raw data* con aplicaciones como Google Refine o GRDDL para datos en XML. Efectuada la transformación se requiere una nueva fase de limpieza de datos por fases: reevaluación de las condiciones de acceso HTTP, errores con la nomenclatura de términos en vocabularios etc. Importante también es la valoración y elección de los métodos de almacenamiento de los datos transformados y la conversión de los metadatos a formatos legibles por máquina. La gestión de los metadatos de los datasets es una parte fundamental del proceso pues establece la visión global del proceso indicando la propiedad de los datos su procedencia y ayudando a controlar su calidad, actualizarlo y preservarlo.
4. Vinculación de los datos: se encuentra en los fundamentos del paradigma Linked Data. El verdadero valor de los datos se manifiesta en tanto que vinculados a otros datasets que aumentan su valor, visibilidad y calidad. El vinculado supone la agregación de datos que en conjunto ofrece nuevas posibilidades de mostrar conocimiento especialmente cuando se vinculan datasets de calidad. Tan importante como el vinculado es el mantenimiento de esos enlaces entre datasets, tarea que debe ser incluida en la gestión de los datos como prioritaria. Cabe la posibilidad de establecer vínculos enriquecedores de modo automático, con herramientas como SILK o SINDICE. Los modos manuales son también posibles a

través de *owl:sameAs*, *rdfs:seeAlso*, *rdfs:subClassOf* o *rdfs:subPropertyOf*, *foaf:knows*, y la localización de conjuntos de datos relacionados utilizando SPARQL. Es recomendable documentar el dataset con una descripción VoID. Hay tres pasos fundamentales a la hora de mapear datos: identificar los datasets candidatos al vinculado, descubrir los datos enlazables, validar las relaciones establecidas.

5. Enriquecimiento de los datos: Los datos en bruto, sin interacción alguna pueden ser enriquecidos mediante diversas técnicas. Auer (2013) afirma que el enriquecimiento a estos efectos es un proceso que aumenta la expresividad y la semántica de una base del conocimiento dada. Este enriquecimiento se consigue mediante la adición o refinado de axiomas terminológicos. El autor propone un enriquecimiento gradual de los datos en vez de la preparación de una base ontológica completa desde el principio, esto coincide con la idea de Berners-Lee de publicar *raw data now* y persigue, dada una base de conocimiento, su análisis para mejorar el *schema* que la articula. Los metadatos descriptivos de los datasets son otra fuente de enriquecimiento potencial: metadatos de preservación digital, de licenciamiento, de procedencia, etc.
6. Publicación de los datos: Podemos estructurar el proceso de publicación en tres partes diferenciadas:
 - a. Publicación del conjunto de datos en un *triplestore*, mediante herramientas como Virtuoso, Sesame, etc.
 - b. Publicación de metadatos: tal y como se dijo más arriba, descripciones de datasets con VoID que deben publicarse para mejorar la visibilidad
 - c. Mejorar el descubrimiento de los datos: utilizar *sitemaps* para buscadores, registrar los datos en repositorios de datos (CKAN).
 - d. Proveer accesibilidad múltiple a los datos, preparando una interfaz humana para la búsqueda y la navegación de los datos, a ser posible a través de un navegador semántico, posibilitando la descarga de los datasets en los formatos más interoperables y facilitando el acceso a un *SPARQL endpoint*.
7. Calidad, control de los datos y preservación de la calidad: se define la calidad de los datos como su aptitud para ser utilizados en un determinado contexto y que se caracteriza por su integridad, suficiencia de la representación semántica o el grado de “legibilidad” de los datos. Por otro lado no se consideran datos de calidad, aquellos que están incompletos, inexactos, son inconsistentes en la representación o con sintaxis inválida. Todo ello se comprueba con los sistemas de validación de datos pertinentes y se controla en una tarea continua que abarca todo el proceso. Al igual que otros objetos digitales, sería

recomendable crear un plan de gestión de los datos que controlará el valor de los datos y el mantenimiento de las condiciones de accesibilidad a los mismos; este proceso de control debería verificar incluso la no publicación de datos, la revisión de enlaces o la migración de tecnologías, etc.

8. Consumo de los datos: estamos ante un punto fundamental del ciclo de los datos vinculados. La generación de datasets bajo tecnologías Linked Data no tendría sentido sin el consumo o uso de los datos, ya sea libre de cargas o no. Como se ha dicho antes, ofrecer puntos únicos de acceso, con posibilidades de búsqueda, navegación y descarga es un requisito necesario tanto de la publicación como del consumo de los datos. Una buena estrategia conlleva aplicar los siguientes puntos:
 - a. El objetivo final de la publicación debe tener en cuenta las necesidades externas. No se consumen los datos carentes de interés.
 - b. La revisión de casos de uso relevante puede ayudar al diseño de estrategias coherentes de consumo.
 - c. Reevaluar las licencias de uso: un consumo deficiente puede estar fundamentado en políticas de licencia inadecuadas.
 - d. Desarrollar seguimientos de consumo para verificar la utilización concreta y parcial de los datos.
 - e. Controlar el alineamiento de vocabularios y la actualización de los datos, a ser posible de modo desatendido.
 - f. Establecer políticas de promoción de los datasets, si es el caso.

3 LINKED OPEN DATA – GLAM (GALLERIES, LIBRARIES, ARCHIVES, MUSEUMS)

No son pocos los que afirman que las bibliotecas y, por extensión, el resto de las instituciones del patrimonio cultural, están inmersas en un profundo proceso de cambio. Las tecnologías de la información están sustrayendo de manera paulatina, el control que antaño estos centros tenían sobre la información en todas sus facetas. Ahora, estos servicios se enfrentan, en nuestra opinión, a un cambio tan radical que amenaza con transformar su propia naturaleza, una adaptación obligatoria hacia un contexto donde las bibliotecas no tienen un lugar preeminente y donde la información se difunde, se utiliza y se comparte al margen de ellas.

Se podría comparar la situación con la de las antiguas bibliotecas medievales, donde se preservaba la riqueza cultural del mundo de puertas para adentro. La llegada del Renacimiento promovió un cambio de paradigma, los libros salieron de su seguro encierro y surgió la necesidad de divulgar extensamente esta información. Las consecuencias son conocidas, una evolución cultural, técnica y artística sin precedentes. Quizás ahora, las bibliotecas, están ante la misma situación, abrir los silos de datos provoca no pocos recelos, pero cada vez hay menos dudas acerca de que las bibliotecas deben adherirse a este movimiento aperturista y colaborativo y participar en él con sus mejores armas: la capacidad de aprender, de adaptarse, de colaborar, de ordenar y de preservar. (Por comodidad expositiva, se pretende que el concepto “biblioteca” haga referencia al mundo de las instituciones de la cultura: museos, centros culturales, archivos, y cualquier tipo de institución análoga).

Linked Data es el vehículo que transporta este nuevo cambio de paradigma, y por eso es interesante conocer cuál es el estado de la situación.

3.1 LINKED DATA Y LAS INSTITUCIONES DEL PATRIMONIO CULTURAL

3.1.1 LIBRARY LINKED DATA INCUBATOR GROUP FINAL REPORT

Linked Data ofrece a las bibliotecas una serie de importantes ventajas en la gestión de los datos: la posibilidad de compartirlos, de hacerlos reutilizables y por tanto favorecer el enriquecimiento potencial de la información al hacer interconectables los sets de datos. Estas posibilidades de interconexión favorecen el trabajo colaborativo entre bibliotecas, la recombinación de conocimiento y el desarrollo de nuevos e innovadores servicios. Las bibliotecas generan metadatos de gran calidad y su publicación bajo Linked Data hace que mejore la accesibilidad de

esos productos de la actividad bibliotecaria, lo que provoca reconocimiento del trabajo y esfuerzo interno de los centros en el exterior.

Compartir datos permite la creación de una red global de conocimiento, que desde el punto de vista de la biblioteca, propicia la construcción de una plataforma en red de ricos metadatos plenamente accesibles y utilizables por cualquiera, haciendo disminuir los trabajos y esfuerzos individuales y redundantes, que con frecuencia consumen los recursos de los centros de información. Linked Data pues, permite la presentación de la información como un todo continuo, con nodos de datos estructurados en RDF e identificables mediante IRIs, conectando contenidos genuinamente bibliotecarios con contextos informativos ajenos que generan un nuevo entorno de conocimiento compartido. (Baker et al., 2011).

En el contexto de las bibliotecas universitarias, específica y estatutariamente dedicadas al apoyo a la investigación, la vinculación de datos científicos o provenientes de la investigación, promoverá un marco de conocimiento que podrá ser aprovechado por otros investigadores desde diferentes puntos de vista: optimizar sus trabajos, utilizar esa información en áreas diferentes a la original como complemento a su propia investigación, refutar o acreditar experimentos asentados, etc. Todo ello favorece la transparencia de la investigación científica, la no reduplicación de esfuerzos y recursos, y la aparición de nuevos y enriquecidos soportes de la investigación, como las publicaciones mejoradas (*Enhanced publications*) (Vanderfeesten, 2012). Aquí las bibliotecas también tienen algo que hacer.

La capacidad de descripción de contenidos, utilización de metadatos y vocabularios, y de preservación de materiales de las bibliotecas puede verse completada con el esfuerzo de otros centros de información u organizaciones públicas o privadas, cuya colaboración puede ayudar a suplir carencias técnicas o presupuestarias. Además, la generalización de los datos vinculados ofrecerá la posibilidad de colaborar con un amplio elenco de desarrolladores de productos para la información, más acostumbrados a las tecnologías abiertas y no propietarias. La utopía de la “nube de información global” puede ser una realidad a través de datos vinculados en bibliotecas, creándose una plataforma que permitirá a los centros con menores recursos acceder a la información de modo gratuito y sin costes de infraestructura (Baker et al., 2011; Coyle, 2013).

En el ámbito interno de las bibliotecas, los datos vinculados impactarán en la mejora de procesos y servicios, como por ejemplo en la catalogación, en la que la reducción de esfuerzos individuales y la posibilidad de compartir los datos descriptivos con toda la comunidad catalogadora, permitirá dedicar más tiempo a otras tareas y ahorrar recursos. Además, las estructuras de descripción semánticas permanecerán disponibles por largos periodos de tiempo, aunque cambien los formatos, ya que la propia definición de datos estructurados incluye la separación del contenido semántico de la sintaxis o formato, lo que redundará en unas mejores condiciones de preservación de los datos bibliotecarios aportando estabilidad y calidad a este sistema. Estos aspectos son fundamentales, la publicación en Linked Data debe ir aparejada a procedimientos de aseguramiento de la calidad y fiabilidad de contenidos, todo ello con una política de

preservación a largo plazo. La fiabilidad es una de las cuestiones objeto de preocupación por la comunidad de la Web Semántica. (Baker et al., 2011; Crupi, 2013; Guerrini & Possemato, 2013).

3.1.2 SITUACIÓN ACTUAL DE LAS BIBLIOTECAS

Los datos bibliográficos están codificados en lenguajes tradicionalmente utilizados en las bibliotecas, fundamentalmente de tipo textual y junto a códigos identificativos como los del formato MARC. Estos datos bibliotecarios pueden ser compartidos en la actualidad, pero a través de sistemas de intercambio de registros (por ejemplo el protocolo Z39.50) que sólo permiten su utilización dentro de las propias bibliotecas de forma local o consorciada. Este fenómeno es el que la bibliografía define como silos de datos y supone que cada biblioteca se constituye como una reserva de datos sólo compartibles en momentos puntuales y para una utilización a su vez local.

Abrir los datos bibliotecarios genera no pocas dudas, por un lado existe la necesidad de cambiar, de adaptarse a los retos que la Web de datos supone, por otro están las dificultades de eliminar sistemas muy arraigados, extendidos y fiables como el formato MARC y modos de hacer que se enfrentan a nuevos procesos más flexibles y colaborativos. Otro problema es que la comunidad bibliotecaria depende en exceso de los sistemas de gestión de datos propietarios, los cuales son provistos por empresas y grupos comerciales del sector que temen que la reconversión a datos vinculados pueda resultar una amenaza a sus modelos de negocio (Baker et al., 2011; Crupi, 2013; Peset, Ferrer-Sapena, & Subirats-Coll, 2011).

La comunidad LODGLAM (Linked Open Data-Galleries, Libraries, Archives, Museums), aunque quizás de un modo algo desordenado, se ha mostrado como una de las más activas en la aportación de recursos bibliotecarios mediante datos vinculados. La mayoría de las grandes bibliotecas nacionales han apostado por servicios Linked Data con diferente nivel de compromiso. De los datos publicados hasta ahora, la mayoría son vocabularios de valores, como las listas de encabezamientos de materia, o conjuntos de elementos de metadatos como los vocabularios RDA (recientemente actualizados). Son menores las publicaciones de datos puramente bibliográficos, como bibliografías, citas, etc., aunque OCLC ha publicado recientemente más de 200 millones de registros bibliográficos en Linked Data.

Existen otros puntos de debate de considerable importancia: la preocupación creciente por la preservación de los set de datos manteniendo los niveles de calidad, el desarrollo de tecnologías que permitan una suficiente automatización de los procesos y que eviten la complejidad primaria de Linked Data, la carencia de competencias suficientes entre el staff bibliotecario para acometer o impulsar proyectos sobre datos vinculados, etc. Las cuestiones legales no son fáciles de definir, no es sencillo atribuir la propiedad de los registros catalográficos integrados por producciones locales o agregaciones de otros catálogos.

El enfoque participativo de la bibliotecas en la Web de datos puede definirse bajo las siguientes premisas: (Baker et al., 2011; Coyle, 2013):

1. Aceptar definitivamente el nuevo marco en el que se desenvuelve la información, haciendo que las bibliotecas publiquen, por defecto, sus datos bibliotecarios en Linked Data asegurando su interoperabilidad y reutilización.
2. Disponer libremente de esos datos para quienes busquen la información a través de licencias abiertas y para el resto de supuestos legales, llevando a cabo acuerdos con los propietarios de los datos que permitan la publicación.
3. Identificar los sets de datos que se vayan a publicar maximizando el aprovechamiento al menor coste posible y buscando proyectos no excesivamente complejos.
4. Mejorar la comunicación entre los bibliotecarios y los miembros de la comunidad semántica, fundamentalmente en los siguientes puntos:
 - a. Ampliar los esquemas de expresión semántica a las necesidades de descripción de los datos de bibliotecas.
 - b. Procurar que los datos de bibliotecas vinculados ofrezcan el máximo de utilidad a los consumidores de datos.
 - c. Participar en el desarrollo y modificación de los estándares de la web semántica, colaborando con los organismos internacionales que los desarrollan.
5. Investigar sobre nuevos servicios experimentales poniendo en común los resultados y a ser posible publicándolos.
6. Seguir buenas prácticas y políticas estandarizadas en la gestión de los datos:
 - a. Creación de espacios de nombres en su caso, ya que aportan coherencia, fiabilidad y estabilidad a los sistemas creados de datos vinculados.
 - b. Patrones de IRIs amigables y persistentes, estableciendo pautas de negociación de contenido adecuadas.
 - c. Control de versiones de vocabularios y procurando la extensibilidad a otras organizaciones.
 - d. Traducción de etiquetas y anotaciones a otros idiomas.
7. Definir mapeos entre los set de datos de las bibliotecas y otros conjuntos de datos relevantes mediante la utilización vocabularios estándar.
8. Mantener la calidad de los datos vinculados mediante las tareas adecuadas de gestión y preservación de los datos.
9. Describir los datasets y catálogos de datos a través de esquemas de metadatos adecuados (VoID-DCAT).

Las bibliotecas también podrían liderar los proyectos de transparencia de máxima actualidad entre los diversos organismos públicos. La actividad Open Data podría aprovecharse de las competencias superiores en el manejo de la información por parte de la biblioteca y de herramientas como los metadatos o vocabularios controlados, cuya utilización no está estandarizada en el campo de la divulgación de datos públicos.

La biblioteca puede actuar de plataforma concentradora de todas estas actividades, eliminando duplicidades que en el momento actual sí se están produciendo y mediante esta cooperación de productos y servicios, propiciar una disminución de costes apreciable para la Administración Pública y un salvavidas en forma de suministros presupuestarios extraordinarios vitales para la financiación de servicios y la mejora de las infraestructuras y colecciones para las bibliotecas (Escolano-Rodríguez, 2013) .

3.1.3 BIBLIOTECAS Y AGENTES EXTERNOS DE LA CULTURA

Es interesante remarcar que las ventajas que Linked Data ofrece en el marco de los recursos del Patrimonio Cultural, pueden extenderse en un contexto de cooperación con el sector comercial de la cultura. La propia dinámica del negocio de la información puede propiciar la utilización generalizada de tecnologías semánticas de representación y descripción de la información. Esto permitirá a los proveedores de información y a las bibliotecas utilizar una lengua común que favorezca la interoperabilidad entre ambos y disminuya el consumo de recursos dedicados a la comunicación entre diferentes plataformas y tecnologías.

Es indudable que los metadatos provenientes del sector comercial pueden enriquecer los recursos de las bibliotecas y los estándares que habitualmente utilizan las empresas, pueden aportar un mayor grado de estabilidad en las transacciones con las bibliotecas. Toda esta interacción mejora la armonización entre ambos sectores y promueve la divulgación de los contenidos culturales.

El proyecto Linked Heritage para el enriquecimiento de Europeana y el desarrollo de tecnologías semánticas en la gestión de objetos digitales aborda, en una de sus líneas de trabajo, la colaboración privado-pública en temas de patrimonio cultural, concretamente la reutilización de los estándares y el enriquecimiento e interoperabilidad de los objetos digitales culturales. Sus principales actividades a estos efectos son (Istituto centrale per il catalogo unico delle biblioteche italiane & European Commission, 2014; Martini, 2013):

1. La identificación de estándares y metadatos de la industria cultural y su utilización en el contexto del patrimonio.
2. Ajuste de las tecnologías semánticas utilizadas estableciendo un mapeo entre los esquemas *RDA/Onix for books* (estándar internacional para la representación y comunicación de información sobre productos de la industria del libro en formato

electrónico) y LIDO (esquema museístico utilizado en el marco de los datos vinculados culturales).

3.1.4 CATEGORÍAS DE DATOS BIBLIOTECARIOS

Conviene apuntar brevemente las categorías que estructuran los diversos tipos de datos de bibliotecas. Los conjuntos de datos son colecciones de metadatos estructurados que, en el contexto de las bibliotecas representan al conjunto de descripciones de recursos bibliográficos. Los datasets se componen de entidades especificadas mediante atributos o propiedades (e.g. “título”) y valores que los concretan (e.g. las materias) o texto libre (e.g. “El Quijote”). Los vocabularios de valores sirven para definir elementos. Estos elementos expresan aspectos del objeto y tienen como contenido un valor estructurado en un vocabulario (tesauro, taxonomía, encabezamiento de materia) o un literal. El valor puede (y debe) estar representado por un IRI, lo que permite una fácil reutilización en otro contexto descriptivo. Por ejemplo: el encabezamiento de materia *Semantic network* de la Library of Congress, está identificado por:

<http://id.loc.gov/authorities/subjects/sh92004914>.

Los conjuntos de elementos de metadatos son las propiedades, las clases o los atributos que estructuran las entidades (se asimilan a las propiedades RDF), y que se agrupan en vocabularios como por ejemplo Dublin Core, RDA o SKOS, por ejemplo, la propiedad RDA *is illustrator Of@en* expresa el agente que ilustró la entidad descrita.

Pongamos un ejemplo de descripción bibliográfica sencilla. Se trata de la obra (entidad) de Berners Lee *Weaving the Web* (representada por el IRI de la Bibliografía Nacional de la British Library), especificada por una propiedad del vocabulario Dublin Core que define el tema de la obra y un valor *World Wide Web* (extraído de un vocabulario de valores: los encabezamientos de materia de la Library of Congress). Para la segunda tripleta, se incluye la propiedad RDA *is autor of* y la vinculación con el registro de autoridad de Berners Lee en VIAF. Como se puede observar los diferentes datos han sido sustituidos por IRIs que permiten la identificación unívoca y la fácil su reutilización.

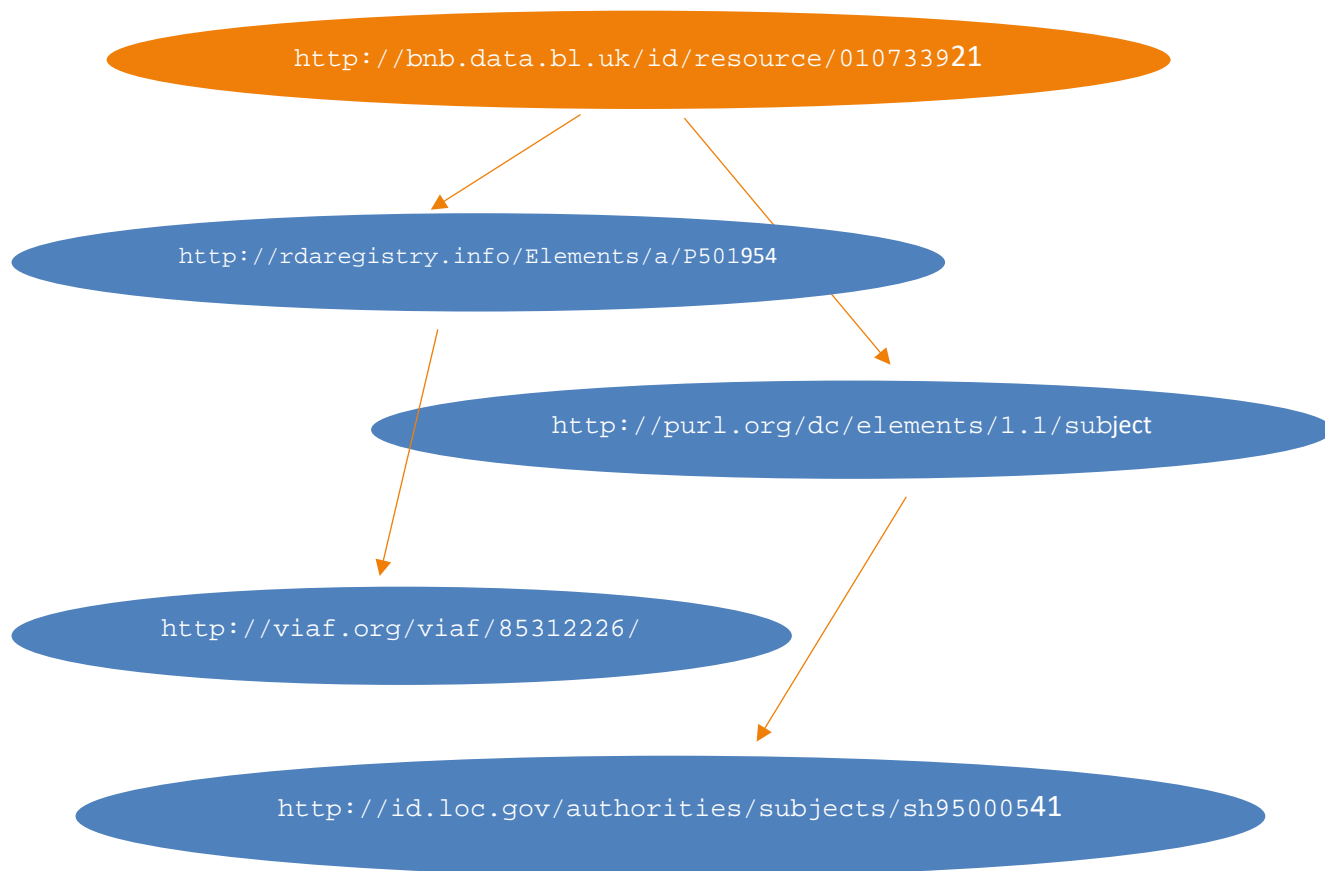


Figura 9 Descripción mediante grafos de la obra *Weaving the Web* de Tim Berners Lee. Fuente: elaboración propia.

3.2 ESTÁNDARES BIBLIOGRÁFICOS DISPONIBLES EN LA WEB SEMÁNTICA

La biblioteca ya dispone de un elenco de herramientas suficientes para comenzar una andadura productiva en la Web Semántica. Las tecnologías Linked Data, al menos a nivel experimental, están presentes en las principales áreas de despliegue de la actividad bibliotecaria. Hay muchos retos que afrontar, una revolución tecnológica que acometer, pero sobre todo, hay que empezar el camino. El mundo de la descripción de recursos, de la recuperación de la información, y de los vocabularios controlados, tienen ya a su disposición multitud de técnicas y herramientas que permiten la conversión de unos formatos robustos, pero creados para entornos locales, en otros en fase beta, pero con unas posibilidades de enriquecimiento de la información insospechadas.

Sean estas tecnologías u otras por venir, lo que es cierto es que las estructuras tradicionales de los sistemas de información, no pueden emplearse con eficiencia en el nuevo entorno. La experiencia y las lecciones aprendidas hasta ahora aconsejan seguir con insistencia y rigurosidad, una serie de pasos: la utilización de estándares de amplio espectro, la descripción de la procedencia en todas sus etapas, la calidad y la veracidad de los metadatos, y la preservación de datos y formatos (E. Mitchell, 2013; Svensson, 2013).

Existe ya un importante número de nuevas iniciativas de lenguajes de descripción de recursos en todos los ámbitos GLAM. La IFLA (International Federation of Library Associations), ha desplegado en formato de esquemas semánticos abiertos, sus principales herramientas bibliográficas: FRBR (Functional Requirements for Bibliographic Records), las FRAD (Functional Requirements for Authority Data) y las FRASAD (Functional Requirements for Subject Authority Data), dan soporte al modelo conceptual de descripción bibliográfica y establecen relaciones (mapeos) con algunos de los más importantes vocabularios del área, también las ISBD ofrecen un rico vocabulario de elementos de metadatos. JSC/COP (Joint Steering Committee for Development of RDA and Co-Publishers) ha puesto a disposición este mismo año, la nueva versión semántica de los vocabularios de RDA, la norma de catalogación candidata a ser un estándar mundial. Otras iniciativas son CIDOC–CRM proveniente del mundo de los museos y en proceso de alineamiento con FRBR a través de la ontología de propósito general cultural FRBROO. También la norma LIDO está disponible como esquema semántico para satisfacer las necesidades de los museos y sus objetos culturales. La comunidad archivística, a su vez también ha desarrollado sus vocabularios semánticos como EAD (Encoded Archival Description), METS (Metadata Encoding and Transmission Standard) y DACS (Cataloging specification Describing Archives). La mayoría de estos estándares ya están publicados en OMR (Open Metadata Registry), mientras se trabaja en el alineamiento entre ellos para mejorar la interoperabilidad en el mundo del patrimonio cultural y se establecen procesos de desarrollo de vocabularios multilingües (Dunsire, Corey, Hillman, & Phipps, 2012).

Como veremos, existen ya importantes propuestas para establecer un marco generalizado de descripción de recursos. Hay que destacar el modelo de datos de Europeana (Biblioteca Digital europea) (National Library of the Netherlands & European Commission, 2014), o de la British Library (British Library, 2014), reflejado en la publicación de su bibliografía en Linked Data. Uno de los desarrollos más importantes es la iniciativa de la Library of Congress BIBFRAME (Bibliographic Framework Initiative) (Library of Congress, 2014a), que no sólo pretende establecer un modelo de catalogación e intercambio de recursos alejado del contexto del formato MARC, sino que también busca ser la base de los futuros sistemas de gestión de bibliotecas. También Schema.org, (un formato de marcado semántico de recursos web a través de microdatos, RDFa, etc.,) ha demostrado su adaptabilidad a los formatos ligeros de descripción embebidos en lenguajes de marcado HTML y está preparando una extensión denominada “Bib”, para definir descripciones de recursos culturales con un mayor nivel de granularidad (Schema.org, 2012). A

primeros de año OCLC (Online Computer Library Centre) (OCLC, 2014a) ha finalizado el proyecto de publicación semántica de sus registros catalográficos en Linked Data con más de 194 millones de registros consultables con una nueva herramienta de recuperación de la información dispuesta al efecto.

En definitiva mucho esfuerzo queda por hacer, por ejemplo apoyar un aumento de las contribuciones semánticas de las bibliotecas, más armonización y estandarización en las publicaciones de datos, más trabajo cooperativo en vez de tantos esfuerzos individuales, sobre todo a nivel de las grandes instituciones de la cultura. La razón del éxito puede estar más en los planteamientos cooperativos para el desarrollo de sistemas comunes (por ejemplo en la catalogación), incluyendo la definición de nuevos vocabularios más estandarizados y la imposición de criterios de calidad rigurosos en la gestión de datos (Dublin Core Metadata Initiative, 2014; Dunsire et al., 2012; E. Mitchell, 2013; Svensson, 2013).

Tabla 4 Diferentes estándares en la estructura de LOD en bibliotecas. Fuente: ALA Report (Mitchell, 2013)

| Concepto | Esquemas |
|----------------------|---|
| Modelado Conceptual | FRBR, FRAD, FRSAD |
| Estructura de datos | RDF, OAI-ORE, XML, EAD, CIDOC-CRM, MARC |
| Contenido | RDA, ISBD, LIDO |
| Intercambio de datos | SPARQL, SRU, OAI-PMH |

En los epígrafes siguientes vamos a hacer una breve referencia a los principales estándares relevantes para bibliotecas y la situación actual de desarrollo semántico de los mismos:

3.2.1 DUBLIN CORE

Las bibliotecas han incluido tradicionalmente entre sus competencias la creación y gestión de vocabularios de calidad. En el contexto Linked Data, los vocabularios, como hemos visto, desempeñan un papel fundamental de apoyo a la descripción de recursos y la estructuración e indización de recursos digitales. En la migración de los vocabularios hacia el espacio de datos vinculados ha desarrollado un papel primordial Dublin Core Metadata Initiative. DCMI promocionó la evolución de los vocabularios hacia su versión semántica haciendo referencia a aspectos nucleares como el alineamiento de vocabularios, la correferencia entre diferentes

esquemas y la mejora de la interoperabilidad general. El alineamiento de vocabularios permite la identificación de equivalencias y distintos tipos de relaciones entre los elementos de metadatos, haciendo que las aplicaciones de datos puedan aprovechar las propiedades de dichos vocabularios incluso fuera del contexto de los propios esquemas.

En el paradigma de datos vinculados, la gestión de vocabularios controlados se hace mucho más compleja, la proliferación de publicaciones y su heterogeneidad dificultan el control de autoridades en las mismas. DCMI promueve el proyecto Vocabulary Community Management que pretende la identificación de buenas prácticas que conduzcan a mejoras en la interoperabilidad, la armonización de los entes publicadores y el multilingüismo; para ello se establecen parámetros para la evaluación de vocabularios, la selección de los mismos, su reutilización y la preservación de vocabularios (Dublin Core Metadata Initiative, 2014; Dunsire et al., 2012).

3.2.2 FRBR Y DATOS VINCULADOS

FRBR es el esquema conceptual para la descripción bibliográfica por antonomasia. Aunque tiene una cierta veteranía, su modelo se adapta con cierta facilidad a las estructuras de datos vinculados, pues su estructura modular encaja bien en el modo distribuido que proporciona Linked Data. Su configuración de elementos en grupos diferenciados y flexibles, permite su utilización por separado, al margen de la estructura monolítica de los registros catalográficos, permitiendo generar estructuras descriptivas independientes, pequeños módulos de metadatos que se pueden utilizar en otros esquemas y vocabularios. (Picco & Ortiz-Repiso, 2012b).

La familia del modelo conceptual FRBR para recursos bibliográficos, se compone también de los modelos FRAD, para registros de autoridad y FRSAD para los datos de materias. Todos estos esquemas conforman un marco de descripción bibliográfica, donde tienen especial cabida las necesidades de los usuarios, que el modelo recoge a modo de tareas. Su expresión como Linked Data se efectúa mediante la enumeración de entidades bibliográficas, las propiedades de dichas entidades y el sistema de relaciones entre los elementos. (Biblioteca Nacional de España, 2014d; Dunsire, 2012; Howarth, 2012).

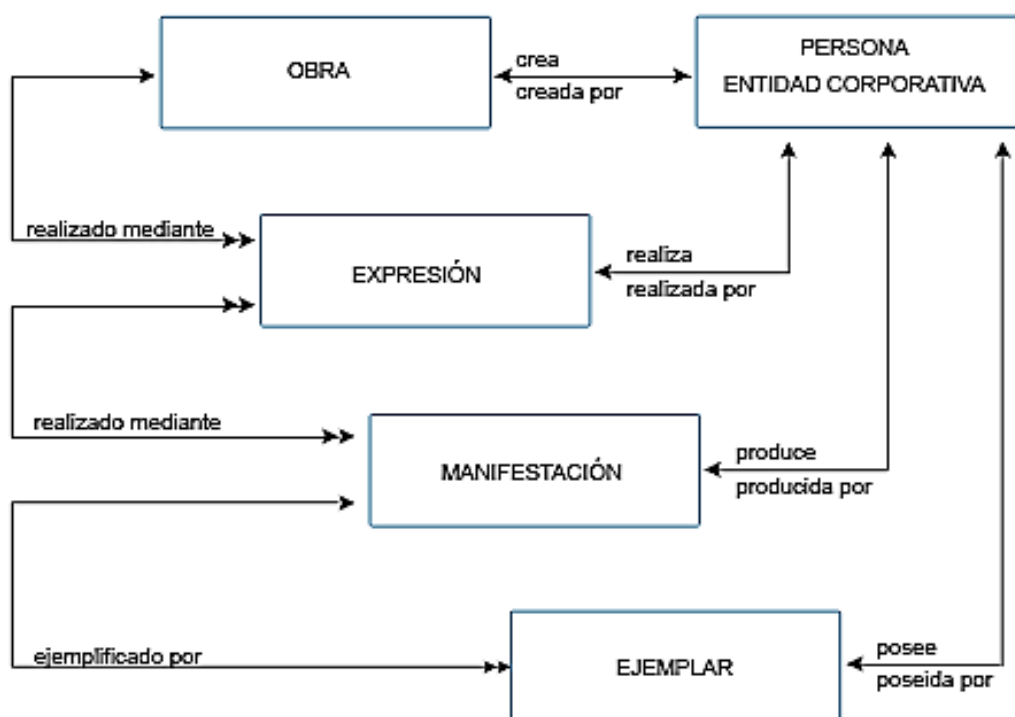


Figura 10 Esquema básico del modelo de entidades y relaciones FRBR. Fuente BNE 2012

FRBR puede cambiar radicalmente el modo de interacción entre las acciones de los usuarios y la información bibliográfica. En este sentido se puede establecer un mapeo de atributos y relaciones y dar soporte con ello a las acciones del usuario, lo que determinaría de modo muy importante la evolución y adaptación de las nuevas normas de catalogación como RDA, en permanente proceso de ajuste. (FRBR define las acciones de usuario como: buscar, identificar, seleccionar, obtener y otras no normativas como: navegar y explorar).

El sistema conceptual que propone, bajo el modelo entidad-relación, ofrece la posibilidad de desagregar los datos de catalogación y vincular estructuras parciales a otros conjuntos de información, conformando nuevas formas no predecibles de descripción. Esto puede ser importante si de lo que hablamos es de la aportación cooperativa de información a través de Linked Data, donde los datos de bibliotecas o vocabularios controlados pueden ayudar a la estructuración de campos complejos como las redes sociales, o donde las etiquetas sociales puedan mejorar la identificación y recuperabilidad de nuestros recursos. La liberación de los datos de bibliotecas puede aumentar lo que Howarth (2012) llama la “biblio-diversidad”, datos de valor que, aprovechando las capacidades de FRBR para desagregar registros y relacionarlos con objetos dentro o fuera de la biblioteca, se ofrecen a la comunidad para su utilización en un contexto de cooperación y de posibilidades tecnológicas como los datos vinculados (International Federation of Library Associations, 2014a; Picco & Ortiz-Repiso, 2012b).

En este marco, Ortiz y Picco (2012) dan un paso más allá y refieren la posibilidad de expresar los registros bibliográficos mediante FRBR separando sus entidades en registros diferentes e identificados unívocamente, lo que permitiría un ahorro considerable a la hora de catalogar manifestaciones de ejemplares, introduciendo únicamente las especificaciones locales de dichas manifestaciones. Además, las posibilidades de vinculación de Linked Data permitirían la utilización selectiva de material conceptual ya publicado a través de relaciones entre registros y sus diferentes niveles de expresión FRBR. Si incluimos en este sistema FRAD y FRSAD las posibilidades se multiplican pues la expresión de las relaciones y datos de autoridad podrían ser también aprovechadas dadas las posibilidades de intercambio y reutilización que los datos vinculados ofrecen (Picco & Ortiz-Repiso, 2012b).

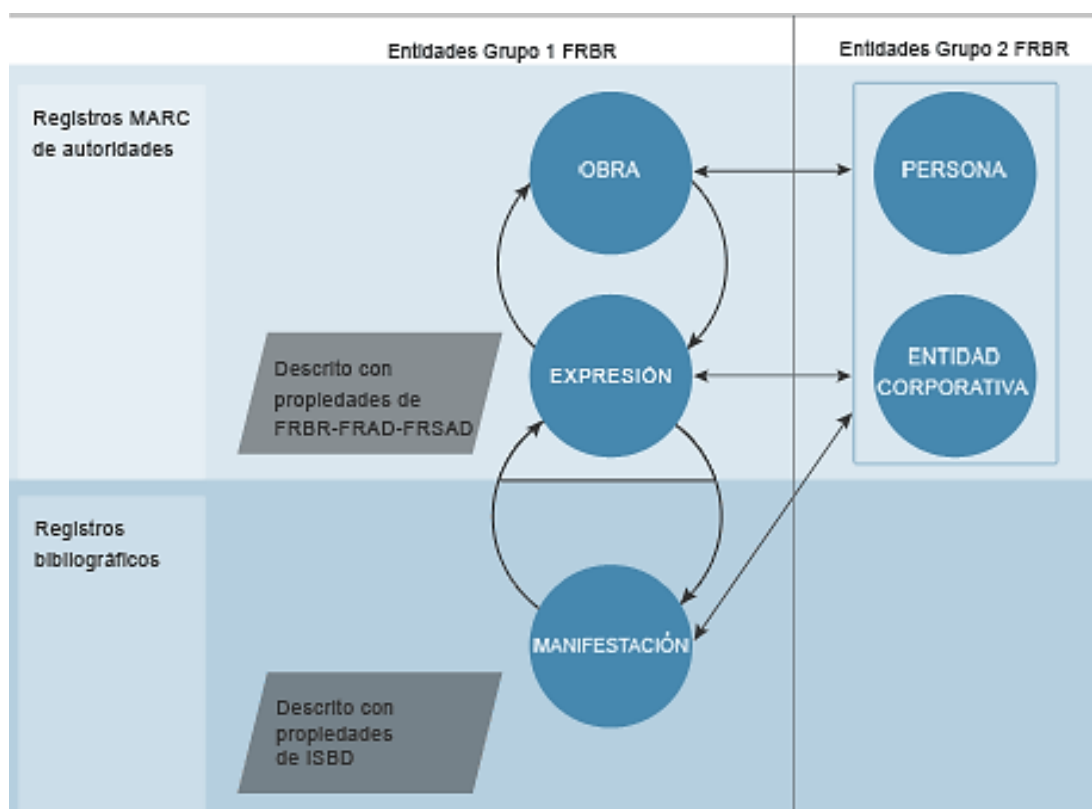


Figura 11 Modelo de RDF de alto nivel para registros bibliográficos. Fuente BNE 2013.

La introducción de FRBR y su familia FRSAD y FRAD en el contexto Linked Data está siendo desarrollada por el FRBR Review Group de la IFLA. Estos estándares ya están publicados en el OMR (Open Metadata Registry) como *namespaces* para su aprovechamiento por la comunidad (National Science Digital Library, 2014). La actualización y armonización de los tres modelos conceptuales ha superado algunas dificultades de definición, como la diferente consideración de

“materia” entre modelos, el ajuste de los “nombres autorizados” o la definición combinada de las tareas de usuario. Otro paso adelante se refiere al modelo FRBRoo (versión de FRBR orientada a objetos y compatible por ello con los recursos del patrimonio cultural), que en concierto con CIDOC CRM (la ontología de descripción documental de los objetos museísticos y del patrimonio cultural), pretende conseguir un modo común de representación del patrimonio cultural, aumentando la interoperabilidad y el intercambio de información entre el mundo de las bibliotecas y los museos (desde 2013 ya dispone de su propio *namespace*). Finalmente se trabaja en el modelo consolidado y único de los cuatro estándares, con su propio espacio de nombres y el ajuste semántico de clases y propiedades. Precisamente las relaciones que se han puesto de manifiesto al intentar vincular elementos de los estándares de la familia FRBR, han puesto de manifiesto las necesidades de ajuste y perfeccionamiento debido a las incoherencias semánticas que se han percibido. (Dunsire, 2012; International Federation of Library Associations, 2014a).

3.2.3 PERSPECTIVAS DEL FORMATO MARC

Es indudable que MARC puede considerarse como una historia de éxito propiciada por su amplia utilización y desarrollo. Este hecho ha provocado que su sustitución sea más compleja de lo esperado, dado el nivel de asentamiento que MARC tiene en todos los sistemas bibliotecarios. MARC está ajustado a las principales normas de catalogación e incluso ha sido adaptado a las nuevas normas tecnológicas, como el esquema de MARC en XML, aunque parece existir un cierto consenso sobre la necesidad de sustituirlo por un nuevo modelo de datos adaptado a las estructuras distribuidas imperantes en el mundo de la información y con un esquema entendible y utilizable por todas las instituciones de la cultura.

El nuevo formato que hay que definir puede quizás completarse con estructuras locales, pero habría que lograr que fueran interoperables para poder hablar de un modelo de intercambio verdaderamente común. Aunque es posible que el modelo deba ser menos ambicioso, no pensando en construir desde cero el futuro estándar de estructuración e intercambio de información, sino simplemente crear un sistema de intercambio de metadatos generalizado desde el que poder crecer (Kroeger, 2013; Moos, 2013).

Parece evidente que la influencia de las bibliotecas en el mundo de la información ha disminuido. La adaptación a nuevos sistemas de intercambio de datos en la Web parece más una cuestión de supervivencia que otra cosa. Hoy la información discurre por canales muy variados y la biblioteca no puede pensar que impondrá su propio sistema de modelado e intercambio de datos. Desde su creación, MARC se ha considerado un obstáculo entre las bibliotecas y el resto de instituciones pertenecientes al ámbito de la información. Los datos bibliotecarios se han mantenido circunscritos a un estricto espacio marcado por los límites definidos por los lenguajes de descripción de recursos y por complejos vocabularios bibliotecarios no fácilmente exportables a otras áreas como los archivos y los museos. En un contexto globalizado como el actual, no parece

posible que la biblioteca siga imponiendo su propio modelo de intercambio de datos. Además MARC tiene algunas debilidades manifestadas en su interacción con el nuevo contexto tecnológico: los nuevos tipos de materiales, su inadecuación de base con las necesidades que requieren los objetos digitales, o incluso su escasa capacidad de representar relaciones entre los recursos. También hay que decir que las resistencias al cambio son fuertes, RDA, por ejemplo es una norma de catalogación que pierde gran parte de su expresividad si se expresa en MARC, no se entiende muy bien cómo se sigue trabajando en la conjunción del formato y la norma. (Kroeger, 2013; McCallum, 2012; Moos, 2013).

3.2.4 ADAPTACIÓN DE ISBD A LINKED OPEN DATA

El estándar ISBD (International Standard Book Description) (IFLA, 2014) tiene también tiene un papel que desarrollar en el ámbito de los datos vinculados. La IFLA considera, dentro de las disposiciones de la versión consolidada (Alcance A.1.2), que se debe mejorar la portabilidad de datos en la Web Semántica propiciando la interoperabilidad del estándar con otras normas. Fruto de esa directriz, la IFLA está desarrollando un marco de cooperación entre los entes participantes en la gestión de la información, y en ese contexto se están desarrollando alineamientos entre principales esquemas del ámbito cultural, (International Federation of Library Associations, 2014b).

El esquema semántico de las ISBD está en pleno proceso de renovación. Por ejemplo, se está trabajando en la versión multilingüe del esquema y reconfigurando los IRIs para transformarlos en IRIs opacos, es decir aquellas que no transmiten ninguna información en un idioma concreto, pues se sustituye el indicador de referencia por un identificador numérico. También se están renovando las asignaciones de etiquetado y nombres de clases y propiedades, intentando ajustar su semántica para su utilización en varios idiomas intentando evitar la ambigüedad que se produce en la utilización en diferentes sistemas. (Escolano-Rodríguez, 2013; International Federation of Library Associations, 2014b; Willer, Dunsire, & Bosancic, 2012).

La publicación del estándar de descripción de recursos bibliográficos ISBD en Linked Data debe considerarse de un modo especial. ISBD no es un formato, es un estándar internacional que está “embebido” en multitud de códigos nacionales de catalogación. Esto hace que la codificación Linked Data de ISBD suponga de alguna manera la extensión de esa expresión fundamental de las normas de catalogación más comunes, aunque este dato no sea muy conocido; incluso FRBR tuvo en consideración para su confección, el estándar ISBD (Escolano-Rodríguez, 2013; Willer et al., 2012).

ISBD siempre ha buscado la interoperabilidad entre bibliotecas y ha sido aséptico desde un punto de vista cultural, lo que ha propiciado su utilización en cualquier contexto. Su alineamiento con RDA a través de un mapeo de sus respectivos *namespaces*, ha sido un paso muy importante en este sentido creándose un perfil de aplicación ISBD compatible con RDA, para que exista

coherencia en la descripción de recursos con ambas normas. El espacio de nombres de ISBD está publicado en OMR (IFLA, 2011) en el marco del Plan Estratégico de la IFLA 2010-2015. Entre los planes de la IFLA para la mejor implicación de ISBD en la web semántica aparece el desarrollo definitivo del perfil de aplicación de ISBD en DCMI (un perfil de aplicación “AP” es un conjunto de metadatos, políticas de aplicación y guías que definen un modo particular de aplicación, en este caso el esquema RDF/XML ISBD). ISBD-AP tiene como misión la mejora del intercambio de metadatos en la Web, ofreciendo como principal característica, la calidad descriptiva que es capaz de aportar el esquema.

ISBD, también está en la base del desarrollo del formato MARC y a día de hoy está perfectamente alineado con UNIMARC y en menor medida con MARC 21. Este alineamiento puede servir para que a través de ISBD-AP se puedan analizar y transformar esos registros en triples, testando su integridad para comprobar errores internos o falta de elementos requeridos en la descripción. ISBD-AP nos permite dividir registros y utilizarlos parcialmente (del mismo modo que FRBR como ya vimos), facilitando el uso de metadatos por las agencias de intercambio de metadatos, tanto para enriquecer como para mezclar con otros esquemas (International Federation of Library Associations, 2014b; Willer et al., 2012).

Las especificaciones de modelado del esquema ISBD, admiten posibilidades bidireccionales de relación entre atributos expresadas estas mediante etiquetas que indican el sentido de la relación (*has title proper*, *is title proper of*). Cada clase en el esquema ISBD va indicada con un número y la letra C y las propiedades llevan la letra P delante del número. Cada área ISBD es un *aggregated statement* (especie de declaración compuesta por un conjunto de metadatos dispuestos habitualmente según un orden y un régimen de obligatoriedad y repetitividad de los elementos); estas declaraciones especiales se especifican en RDF como un “SES” (*syntax encoding scheme*), recurso que proviene de las RDA y que permite una especial configuración sintáctica de algunos elementos determinados de la descripción. Este mecanismo soporta diferentes niveles de granularidad en la descripción, los IRIs se asignan a los esquemas de codificación de cada área, y puede ser referenciados en las Dublin Core Application Profile de ISBD como módulos coherentes. (Willer et al., 2012).

Para la integración con otros esquemas, el proyecto define sus clases y propiedades en un nivel jerárquico inferior a aquellos, dejando el nivel más alto de la descripción otros *namespaces* como por ejemplo Dublin Core, lo que permite la utilización de sus elementos como subpropiedades o subclases, lo que le convierte en un instrumento útil para una especificación más detallada de los registros (Willer et al., 2012).

3.2.5 RDA Y LINKED DATA

3.2.5.1 Aspectos generales de RDA en el marco de los datos vinculados.

RDA (Resource Description & Access) es un nuevo código de catalogación (sustituto de las AACR2) creado y gestionado por el Joint Steering Committee for Development of RDA (JSC) y que se encuentra ya en uso en algunas de las principales bibliotecas del mundo como la Library of Congress, la British Library o la Deutsche National Bibliothek. Durante su elaboración se han tenido en cuenta, a efectos del tema que aquí se trata, los diferentes esquemas de metadatos de las comunidades del patrimonio cultural, procurando que los elementos RDA tengan un nivel previo de alineación suficiente con los mismos.

Los vocabularios de RDA también están publicados en OMR, aunque a primeros de año se han puesto a disposición de la comunidad en un registro específico y remozado (Joint Steering Committee for Development of RDA, 2014). Los vocabularios RDA publicados ofrecen elementos que proporcionan posibilidades de descripción y acceso a recursos mejoradas. Su identificación mediante IRIs permite a los sets de elementos de RDA convertirse en una referencia de codificación de primer nivel (Picco & Ortiz-Repiso, 2012a; Tillet, 2013). En las nuevas versiones de los esquemas se ofrecen características que favorecen la integración con FRBR, y por compatibilidad, se mantienen versiones para sistemas de catalogación que no utilizan todavía el modelado conceptual. Desde un punto de vista técnico, se han equilibrado los IRIs legibles por máquina y por humanos para facilitar su reutilización y también se está trabajando en la mejora de la interoperabilidad con nuevos proyectos de alineamiento. La distribución completa de vocabularios RDA, disponibles desde primeros de año, es la siguiente (American Library Association, 2014; Joint Steering Committee for Development of RDA, 2014):

1. *Classes*
2. *Work properties*
3. *Expression properties*
4. *Manifestation properties*
5. *Item properties*
6. *Agent properties*
7. *Unconstrained properties*

Los elementos RDA pueden proveer instrucciones especiales, con diferentes niveles de granularidad en la descripción, esto permite un nivel óptimo de adaptabilidad a las necesidades de catalogación, característica fundamental dada la multiplicidad de materiales, formatos y soportes de los recursos actuales. Lo que nos ofrece RDA es un esquema de codificación íntegro, con etiquetas bien establecidas y un sistema eficiente de gran potencial para indicar relaciones. Como es habitual en los vocabularios de propiedades, RDA puede vincularse a diferentes niveles

jerárquicos, por ejemplo como subpropiedades de los elementos que requieren menor granularidad (sistema análogo al definido antes para ISBD). (Danskin, 2013).

RDA tiene características que conjugan perfectamente con el ecosistema Linked Data; por un lado vincula la descripción de recursos a las tareas de los usuarios y por otro a las entidades específicas que contienen los recursos (influencia FRBR). Sus vocabularios son directamente aplicables a la catalogación a falta únicamente de interfaces y sistemas bibliotecarios adaptados, aportando un importante ahorro de costes, una descarga en las tareas, la posibilidad de catalogación multilingüe y cooperación armonizada con otros agentes de la cultura (Tillet, 2013). Todos los objetos culturales de los que dispone la biblioteca pueden ser codificados con sus principales características, no sólo los bibliográficos, esta codificación puede ser expresada en términos Linked Data y mediante el procesamiento automático, exponer esos objetos en la Web, donde podrán ser recuperados por los usuarios para enriquecer sus contenidos y mejorar la visualización de esos recursos. (Picco & Ortiz-Repiso, 2012a; Tillet, 2013)

El JSC RDA Task Group, coordinadamente con la IFLA tiene entre sus programas de trabajo el alineamiento total de ISBD y RDA, el programa *Machine-Actionable Data Elements* para mejorar el proceso automático de los datos creados y la representación en RDF del sistema de relaciones RDA. Como propuesta de futuro DCMI y JSC trabajan en un proyecto todavía embrionario, en el desarrollo de un perfil de aplicación de RDA y Dublin Core basado en los estándares de la familia FRBR (Joint Steering Committee for Development of RDA, 2014).

3.2.5.2 Descripción mediante RDA de un recurso de la biblioteca UPM

Algunos IRIs referentes a *namespaces* del sistema bibliotecario BUPM se han creado a modo de ejemplo para poder efectuar la descripción.

```
@prefix upmcat: <http://www.upm.es/biblioteca/cat#> .
@prefix rdaa: <http://rdaregistry.info/Elements/a/> .
@prefix rdae: <http://rdaregistry.info/Elements/e/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix rdam: <http://rdaregistry.info/Elements/m/> .
@prefix rdaw: <http://rdaregistry.info/Elements/w/> .
```

```
# Descripción bibliográfica del libro: "Example: Maxwell's handbook for RDA:
resource description and access using MARC 21" Fuente:
http://www.rdaregistry.info/#
```

```
<http://www.upm.es/biblioteca/cat/itembib/6500078586> a
dcterms:BibliographicResource.
```

```
#Agent properties, <http://rdaregistry.info/Elements/a/>#
```

```
upmcat:Agent rdaa:P50094
"http://viaf.org/viaf/266378769/#Maxwell,_Robert_L.,_1957-...." ;
#identifierForThePerson#
```

rdaa:P50117 "Maxwell, Robert L." ; #preferredNameForThePerson#
rdaa:P50121 "1941" . #dateOfBirth#

#Expression properties <<http://rdaregistry.info/Elements/e/>>#

upmcat:Expression rdae:P20001
[<http://rdvocab.info/termList/RDAContentType/1020>
skos:prefLabel "text"] .
upmcat:Expression rdae:P20206 "Incluye tabla de contenidos, índice de
tablas, índice de figuras y bibliografía" ; #supplementaryContent#
rdae:P20231 upmcat:Expression . #workExpressed#

#Manifestation properties <<http://rdaregistry.info/Elements/m/>>#

umpcat:Manifestation rdam:P30001
[<http://rdvocab.info/termList/RDACarrierType/1049>
skos:prefLabel "volume"] ;
rdam:P30002
[<http://rdvocab.info/termList/RDAMediaType/1007>
skos:prefLabel "unmediated"] ;
rdam:P30003
[<http://rdvocab.info/termList/ModeIssue/1001>
skos:prefLabel "single unit"] ;
rdam:P30004 "ISBN 978-0-8389-1172-3" ;
#identifierForTheManifestation#
rdam:P30011 "2013" ; #dateOfPublication#
rdam:P30088 "Chicago" ; #placeOfPublication#
rdam:P30135 upmcat:Work ; #workManifested#
rdam:P30139 upmcat:Expression ; #expressionManifested#
rdam:P30141 "<http://www.ala.org/>" ; #contactInformation#
rdam:P30156 "Maxwell's handbook for RDA : resource description and
access using MARC 21" ; #titleProper#
rdam:P30169 "25 cm" ; #dimensions#
rdam:P30176 "American Library Association" ; #publishersName#
rdam:P30181 "910 p." ; #extentOfText#

#Work properties <<http://rdaregistry.info/Elements/w/>>#

rdaw:P10002
"http://viaf.org/viaf/266378769/#Maxwell,_Robert_L.,_1957-...." ;
#identifierForTheWork#
rdaw:P10061 upmcat:Agent ; #author#
rdaw:P10088 "Maxwell's handbook for RDA : resource description and
access using MARC 21" . #titleOfTheWork#

3.2.6 SCHEMA ORG

Schema.org está constituido por una serie de vocabularios desarrollados por las grandes compañías proveedoras de servicios de información comerciales, como Google, Microsoft o Yahoo, para potenciar la semántica de los contenidos y mejorar las búsquedas de información en la Web. La utilización de Schema.org en general posibilita la creación de descripciones enriquecidas de recursos heterogéneos en la Web, estableciendo un lenguaje común para los diversos agentes que participan en la misma: bibliotecas, proveedores de contenidos, publicadores, buscadores, etc. Ofrece un marco de rápida y fácil aplicación, que permite el marcado masivo y las actualizaciones de datos fácilmente automatizables, todo ello bajo la infraestructura ofrecida por la OCLC para los datos y el catálogo WorldCat.org (Fons, Penka, & Wallis, 2012; Schema.org, 2012)

Schema.org está compuesto por tipos de elementos y sus propiedades. Su flexibilidad se apoya en la multiplicidad de vocabularios específicos como: CreativeWork, Book, Movie, Event, Organization, Person, Place, Product, etc. Estos vocabularios se serializan dentro del código HTML, habitualmente en microdatos, RDFa o JSON-LD, semantizando sus contenidos y permitiendo por ello el proceso automático de la información (Fons et al., 2012; Schema.org, 2012). El esquema ha sido recientemente mejorado mediante la incorporación de la ontología de comercio electrónico *GoodRelations* que también está respaldada por los gigantes Yahoo y Google (Godby, 2013). Schema.org está desarrollando una extensión denominada *Schema Bib Extended*, que permite la descripción específica de materiales bibliográficos.

Schema.org propone una infraestructura que en referencia al espacio de las bibliotecas se ajusta al siguiente modelo:

1. Un esquema con varios vocabularios orientados a la descripción semántica de la web, con la posibilidad de enriquecimiento semántico con otros vocabularios.
2. La extensión en desarrollo *Bib Extended*, que permitirá una mejor descripción de los contenidos bibliotecarios. Un acuerdo global de todos los agentes facilitaría la extensión de Linked Data de modo amplio a las bibliotecas.
3. Acceso a datos bajo múltiples serialización RDF.
4. Mejora del servicio de descubrimiento del catálogo Worldcat.org.
5. Colección de enlaces a recursos como VIAF o FAST accesibles e identificables para su reutilización. Adición de nuevos enlaces de recursos autorizados.
6. OCLC acaba de publicar cerca de 200 millones de registros en Linked Data a través de Schema.org, integrando recursos de aplicación semántica como como FAST (*Faceted Application of Subject Terminology*), VIAF (*Virtual International Authority File*), LCSH (*Library of Congress Subject Headings*) y el sistema de clasificación DEWEY.

La convergencia entre los intereses y posibilidades del entorno clásico LOD y Schema.org ofrece ciertas dificultades. La abstracción propia del modelo FRBR para recursos bibliográficos no es muy comprensible para usuarios que pretenden marcar semánticamente recursos diversos en la Web de un modo rápido y simple. La integración de la extensión de biblioteca de Schema.org no debe aportar complejidad extra a un esquema cuya filosofía establece un uso generalizado y fácil de aplicar. Por ello la extensión de biblioteca aparece más como un complemento que se utilice para descripciones más estrictas, en las que se deben especificar conceptos tan abstractos como Obra, Expresión, Manifestación, etc., o describir diferentes formatos de publicaciones, o traducciones de un mismo trabajo.

Existe un proceso de alineamiento en marcha entre Schema.org y BIBFRAME cuyo objetivo es conseguir un formato robusto y sencillo de utilizar y que aproveche las mejores características de ambas plataformas descriptivas. En ese contexto se están promoviendo algunos cambios en la sintaxis de Schema.org para la integración con BIBFRAME, al menos para las descripciones menos exigentes. Se trata de la aplicación de propiedades expresivas de jerarquías FRBR, como *hasInstance* o *isInstanceOf*, que permiten un más estrecho alineamiento con el vocabulario propio de BIBFRAME (Godby, 2013). Efectivamente, BIBFRAME está diseñado para la utilización por catalogadores que efectúan descripciones bibliográficas de nivel bibliotecario y como formato de intercambio de estos registros. La conjunción de ambos esquemas puede redundar en una mayor recuperabilidad de los recursos por buscadores web (Schema.org) y una mayor consistencia, mejor gestión y curación de los recursos bibliográficos (BIBFRAME).

Un ejemplo de descripción en Schema.org a través de microdatos puede ser el siguiente, también las IRIs BUPM se han creado a modo de ejemplo:

```
<div itemscope itemtype="http://schema.org/Book" itemid="
http://upm.es/library/cat/item/6000084373">

  Title: <span itemprop="name">Web semántica y sistemas de información documental
</span><br/>

  Author: <a itemprop="author" href="http://viaf.org/viaf/39656622/">Lluís
Codina</a><br/>

  ISBN: <span itemprop="isbn">9788497044608 </span><br/>

  Publisher: <span itemprop="publisher">Trea</span><br/>

  Genre: <span itemprop="genre">Web semántica</span><br/>

  Date Published: <span itemprop="datePublished">2009</span><br/>

  Pages: <span itemprop="numberOfPages">297</span><br/>

  Alternate description: <a itemprop="sameAs"
href="http://www.worldcat.org/oclc/464748065">WorldCat</a><br/>
</div>
```

3.2.7 LINKED OPEN DATA ENABLED BIBLIOGRAPHIC DATA 2.0 LOD-BD

La FAO (Food and Agriculture Organization of the United Nations) a través de la AIMS (Agricultural Information Management Standards) ha promovido el proyecto hacia la definición de procesos Linked Data para la descripción de datos bibliográficos. Con este fin utiliza un conjunto de vocabularios estándar que describen en todo su ámbito, recursos bibliográficos de muy diferente índole y además, apoya el proceso mediante la presentación de árboles de decisión que ayudan a la especificación de las propiedades de los recursos. LOD-BD pretende ser una guía eminentemente práctica, donde se abordan detalladamente las estrategias y procesos para la descripción de los recursos bibliográficos en el contexto de Linked Open Data. Para ello utiliza estándares que aseguran la interoperabilidad con el objetivo de compartir la información.

Concretamente el modelo conceptual LOD-BD (un sistema FRBR reconvertido y adaptado) define tres tipos de entidades: recursos, agentes y temas. El recurso es el objeto bibliográfico, archivístico o museístico; el agente es el ente creador, ya personal ya colectivo y el tema o materia expresa los diversos contenidos que especifican los recursos bibliográficos. Los recursos son las entidades básicas del modelo con relaciones principales establecidas con los agentes y temas.

Las relaciones principales se establecen entre instancias de una misma entidad o entre entidades. El control de autoridades es una parte importante del modelo: los

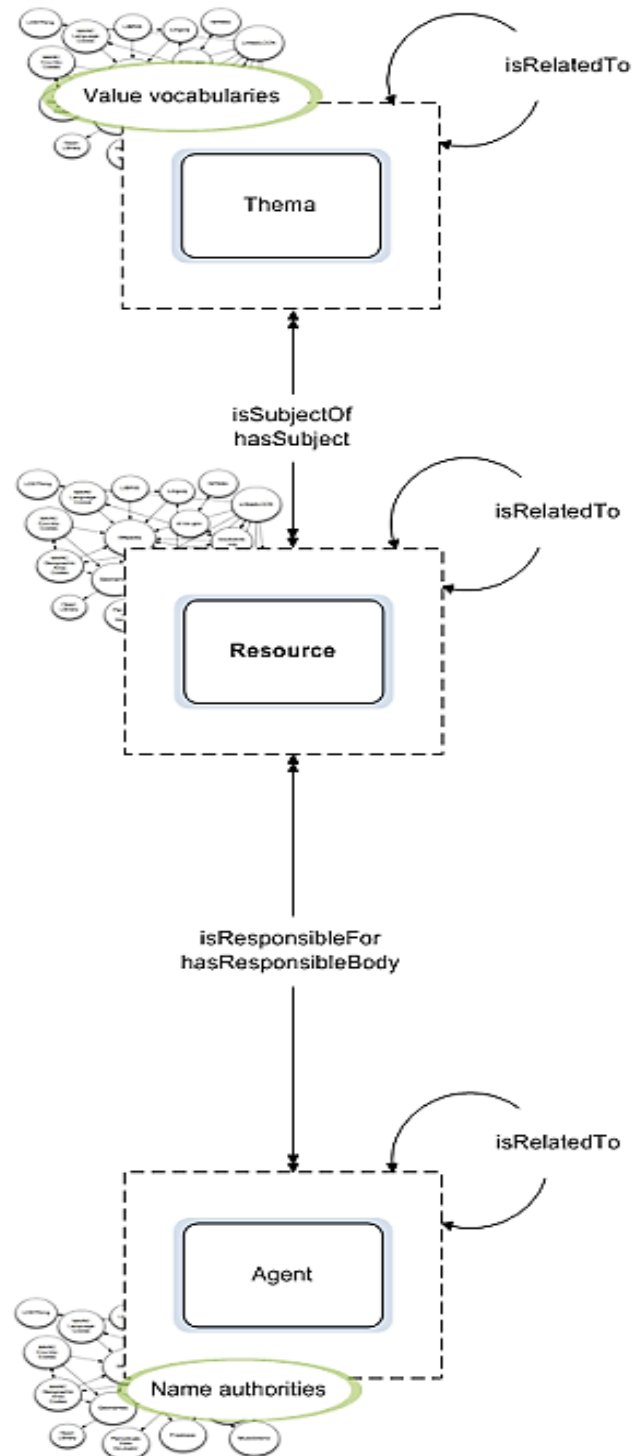


Figura 12 Modelo de datos LOD-BD 2.0. Fuente: Agricultural Information Managements Standards 2012.

agentes se gestionan mediante archivos de autoridad ya disponibles en la nube LOD (por ejemplo VIAF); los conceptos, temas y lugares se controlan mediante vocabularios de valor también en la red (LCSH, RAMEAU, etc.).

Los recursos son especificados por nueve grupos de propiedades comunes y las relaciones entre los recursos por más de 20 propiedades:

1. Título: con las propiedades definitorias de los diversos tipos.
2. Ente responsable como el autor o el editor.
3. Características físicas del recurso.
4. Ubicación física.
5. Materia, en un concepto de significado amplio como palabra clave, término, descriptor, etc.
6. Descripción del contenido: desde resúmenes a notas.
7. Propiedad intelectual definida mediante propiedades que hagan referencia a los derechos de autor, derechos de acceso, etc.
8. Uso del recurso: uso público por ejemplo.
9. Relaciones: como se ha referido, reflejan las establecidas entre entidades o instancias de los recursos.

Los *namespaces* utilizados para las descripciones son Dublin Core Metadata Element Set (dataset básico), DCMI terms (que incluye todos los metadatos de DCMI, incluidas clases, propiedades, esquemas de codificación de vocabularios y esquemas de codificación de sintaxis), BIBO (ontología para la descripción de recursos bibliográficos), AGLS Metadata Standard (estándar que proporciona metadatos para la mejora de la visibilidad e interoperabilidad de los recursos de información y servicios en línea), Eprints (esquemas de metadatos para la descripción de eprints y trabajos de investigación) y MARC List for Relators (para la descripción de un nombre y el recurso bibliográfico relacionado).

Todos ellos son estándares muy reconocidos y con suficiente campo para recoger ampliamente la descripción de los recursos. Los procesos de la catalogación se especifican con diagramas de flujo que conforman árboles de decisión que guían en todo momento la descripción, aportando una manera organizada de resolver un problema a la hora de asignar los metadatos (Subirats & Zeng, 2013).

3.2.8 BIBFRAME

3.2.8.1 Norma de descripción bibliográfica y formato de intercambio de la información

BIBFRAME es la propuesta de la Library of Congress para la definición de un nuevo marco de representación e intercambio de la información. Establece un modelo conceptual con cierto ascendente de FRBR (aunque simplificándolo). Pretende diferenciar los contenidos concebidos conceptualmente, de sus posibles y diversas manifestaciones físicas (Obra-manifestación vs Obra-instancia). Por otro lado, establece la completa diferenciación de entidades (identificadas con IRIs), lo que evita la ambigüedad en su reconocimiento permitiendo descubrir y representar las posibles relaciones que se puedan establecer entre los recursos. Este diseño establece una relación bidireccional entre la descripción de contenidos de las bibliotecas y los metadatos que serán expuestos en la Web de datos, de tal modo que la información interna de los centros pueda aparecer como parte de una gran red de datos, es decir una marco específicamente diseñado para datos vinculados con todo lo que ello supone: compartir los datos, enriquecerlos, mejorar la recuperación de la información, etc.

Las características principales que definen el formato BIBFRAME pueden enumerarse como sigue (Library of Congress & Zepheria, 2012):

1. Flexibilidad para la adaptación a los sistemas presentes y futuros de catalogación.
2. Soporte garantizado para nuevas fuentes, tipos de información y soportes.
3. Descentralización de la información bibliotecaria al expresarse a través de la Web.
4. Permite la comunicación bilateral con el entorno exterior a los centros de información, lo que supone una mayor extensión y adopción social de los contenidos y técnicas internas de los centros, a la vez que el contexto exterior influye en el contexto técnico y personal interior.
5. Automatización de procesos descriptivos, lo que facilita al bibliotecario mantenerse al margen de las complejidades técnicas, aunque manteniendo el control intelectual y de creación de contenidos.
6. Aplicabilidad de la plataforma a otros centros del patrimonio cultural, como archivos y museos, lo que supone de facto una mejora sustancial en la intercomunicación y la interoperabilidad derivada de la utilización de un mismo lenguaje y superando con ello las limitaciones del formato MARC.

BIBFRAME ofrece una alternativa de intercambio de la información desde el punto de vista de la descripción bibliográfica. Aceptar su modelo puede acelerar la migración desde MARC, aunque sería deseable que esa transición se produzca a través de un proceso gradual, sin rupturas

innecesarias y a un ritmo que permita la adaptación de la mayoría de la comunidad bibliotecaria. Uno de los principales problemas de MARC, a estos efectos, es su vinculación casi exclusiva con el mundo de las bibliotecas, BIBFRAME abre su modelo compatible con el resto de participantes en el mundo de la información, posibilitando la tan ansiada interoperabilidad real de los datos de todas las instituciones del patrimonio cultural. (Ford, 2012; Kroeger, 2013; McCallum, 2012).

De especial interés es la experiencia y aplicación práctica del estándar BIBFRAME en la Universidad George Washington. Durante este proyecto se han puesto de manifiesto algunos beneficios a corto plazo para los servicios de información. BIBFRAME ha mostrado sus mejores cualidades: un entorno sencillo de trabajo, aprendizaje y desarrollo, y una experiencia de trabajo para el personal que les prepara para futuros retos de más calado y complejidad. La colaboración con otros operadores que experimentaban en el mismo sentido, fundamentalmente los agentes del sector comercial de la información, ha facilitado las mejoras y ha minimizado la comisión de errores comunes. La participación de la comunidad universitaria también ha sido un factor positivo, que ha propiciado un ambiente de cooperación entre los diferentes estamentos universitarios. (Shieh, 2013).

3.2.8.2 Modelo de datos BIBFRAME

EL modelo conceptual de BIBFRAME se basa en una estructura de entidades, atributos y relaciones entre esos elementos sobre FRBR (versión simplificada) y elementos RDA. RDF se encarga de la estructuración de los datos, además de permitir el enriquecimiento mediante diferentes vocabularios controlados y anotaciones como extensiones de los mismos. BIBFRAME ha sido diseñado para ofrecer un marco amplio de utilización donde comunidades diferentes puedan aportar diferentes puntos de vista de un mismo recurso (algo parecido a lo que veremos sobre Europeana y sus proxies). Sus posibilidades de modelado son muy amplias siendo plenamente adaptables a cualquier tipo de serialización (Kroeger, 2013; Library of Congress & Zepharia, 2012; McCallum, 2012). La estructura de datos de BIBFRAME se basa en las siguientes clases:

Creative Work: es un elemento conceptual que representa el trabajo creativo de un autor. Es el punto central de la descripción y alrededor de él aparecen las diferentes instancias materiales de ese recurso. Contiene propiedades y atributos específicos. (Correspondencia con FRBR: Creative Work vs Obra y Expresión).

Instance: consiste en la materialización de un trabajo creativo de cualquier índole, definida por propiedades específicas del recurso material y las relaciones con otros elementos como por ejemplo la publicación o distribución del recurso. Una regla de estructuración es que un trabajo creativo, puede estar materializado por una o varias instancias, mientras que una instancia sólo puede vincularse a un trabajo creativo. (Correspondencia con FRBR: Instancia vs Manifestación y Expresión).

Authority: es un elemento que define un concepto autorizado y relacionado con la obra o la instancia. Sirve para especificar el recurso distinguiéndolo de otros y para potenciar ciertos elementos que permitan una mejor navegabilidad por los recursos. Autoridades a efectos de BIBFRAME son las personas, los lugares, los temas del recurso, etc.

Annotation: recurso que permite una mayor granularidad en la descripción de las clases principales, aportando información relevante sobre la obra o la instancia de un recurso. Algunos ejemplos clarificadores respecto a las obras (*creative works*) pueden ser las revisiones, las tablas de contenido, los resúmenes, etc; respecto a las instancias (*instance*), las imágenes de cubierta, las colecciones localizadas de la biblioteca, la información bibliográfica del autor, los metadatos administrativos, información curada de alta fiabilidad, los datos locales de una biblioteca determinada, la información de procedencia, etc.

El modelo general de descripción de un recurso en BIBFRAME parte de la identificación de una obra o *Creative work* (recurso conceptual) a la cual están vinculadas una o varias instancias o *instances* (representación material ya sea física o electrónica de un recurso). Los metadatos que definen la “obra” son un conjunto de datos vinculados previamente al registro de autoridad como por ejemplo el “Título uniforme”, por ello cada recurso ha de tener por lo menos, una descripción de “obra” y si existe una expresión física o electrónica, una descripción de “instancia”.

La mayoría de clases y subclases pueden llevar anejas unas propiedades generales como `label`, para el etiquetado con un literal; `identifier`, para la asignación de identificadores a los recursos y `authorizedAccessPoint`, para la designación de un punto de acceso según las reglas de catalogación utilizadas. Si están presentes las clases `Authority` y `Annotatiton` en alguna descripción, las propiedades generales que establecen la relación son `hasAuthority` y `hasAnnotation`.

Los tipos de contenido de un recurso se establecen a través de subclases como `Audio`, `Text`, `Dataset`, etc. Para designaciones de otros tipos de elementos específicos como eventos, objetos digitales complejos o piezas de museo es posible la combinación de clases BIBFRAME o acudir a vocabularios específicos.

Las instancias pueden describirse también a través de sus propiedades específicas como `Print` o `Manuscript`. Adicionalmente pueden contener subclases de modos de publicación, como `Monograph` y `Serial`. Se incluyen tres propiedades que son ejemplo de los trabajos de alineamiento con las normas RDA: `contentCategory`, tipo de contenido, `mediaCategory`, o tipo de medio y `carrierCategory` o tipo de soporte.

La información de título aparece en ambas categorías: obra e instancia. El título de “obra” se expresa mediante `WorkTitle` y se refiere al literal que identifica la “obra”, asimilándose según las reglas de catalogación utilizadas con el título preferido, uniforme, etc. El título de “instancia” tiene varias propiedades para especificar facetas del mismo: `TitleValue`, `Subtitle`, `titleSource` etc.

Las propiedades para las instancias hacen referencia a las descripciones típicas de una manifestación de una “obra” (del mismo modo que en FRBR) expresando datos sobre edición, dimensiones, modo de publicación (*edition*, *extent*, *modeOfIssuance*), etc. También puede ser necesario establecer datos descriptivos del proveedor y su relación con la “instancia”, como la producción, la publicación o la distribución; para ello existe la propiedad *providedRole* o *providedStatement*, para transcribir los datos del proveedor al registro.

La clase *Identifier* se expresa mediante un símbolo o cadena de caracteres y se asocia a un recurso para identificarlo. Existe también la propiedad *identifier* que especifica identificadores adicionales que representan unívocamente una instancia. Algunas propiedades de esta clase son: *isbn10*, *isbn13*, *ansi*, *issn*, etc.

Los recursos descritos pueden ser enriquecidos con notas informativas. La propiedad general *note* ofrece información textual sobre el recurso. Existen otras propiedades designadas para supuestos de información específica como *copyNote* (información sobre la copia que se está describiendo), *creditNote* (nota sobre los créditos de los participantes en una producción), etc.

Las declaraciones de materia pueden reflejar dos tipos de información: términos de materia y clasificaciones. Respecto al primero se expresan mediante la clase *Authority*, y la subclase *Topic* pudiendo ser literal o IRIs. Las clasificaciones pueden ser clases o propiedades. *Classification*, como clase, expresa el sistema de organización o codificación utilizado para la asignación de la materia; *clasification* como propiedad indica el número o símbolo concreto asignado según el sistema de clasificación utilizado. Existen gran cantidad de propiedades de clasificación que pretenden describir el contexto de la asignación de materia desde variados puntos de vista.

Respecto a las relaciones entre elementos BIBFRAME, podemos distinguir dos tipos, los cuales pueden ser expresados mediante el vocabulario propio o con otros *namespaces* siempre que sean nombrados en la declaración:

1. Relaciones del agente que expresan el papel de mismo en referencia al recurso como: *creator* o *contributor* y pueden vincularse con las *Authority* de BIBFRAME mediante la propiedad *agent*.
2. Respecto a las relaciones entre dos recursos catalogados existen multitud de propiedades desde más estrictas a más generales:
 - a. Las más generales establecen relaciones entre los principales elementos: obra a obra, instancia a instancia.
 - b. Las generales utilizan propiedades coincidentes con las categorías de relación de RDA, como: *hasEquivalent*, *hasPart*, *accompanies*, *precedes*, *succeeds* o *hasDerivative*.

- c. Las específicas pueden establecer relaciones entre cualquiera de las anteriores como: `supplementTo` (accompanies), `supersedes` (precedes), `continuedBy` (succeeds), and `translationOf` (hasDerivative).

El alineamiento de las relaciones entre BIBFRAME y el conjunto de reglas FRBR y RDA se establece mediante las propiedades `hasExpression` o `expressionof` y para la conexión con instancias BIBFRAME se utiliza `hasInstance` o `instanceof`.

Para las anotaciones las principales propiedades son: `cover art`, `summary`, `review`, `table`

`of contents`, and `holdings`. Las anotaciones generales se efectúan con la propiedad `annotates`. Las informaciones que se pueden incluir mediante el vocabulario de anotaciones abarcan la inclusión de la fecha en la que la declaración fue hecha, quién hizo la declaración y la fuente del contenido de la anotación.

Las principales clases para los fondos bibliográficos son `HeldMaterial` que da información general de los fondos catalogados para un ítem en concreto, por ejemplo las políticas de reproducción. `HeldItem` ofrece información concreta del ítem respecto a su lugar en los fondos: como la signatura o el código de barras (Library of Congress & Zepheria, 2012).

El modelo BIBFRAME utiliza elementos de autoridad que establecen relaciones tanto con la “obra” como con la “instancia”. Estos elementos de autoridad pueden provenir de conjuntos reconocidos de datos de autoridad externos, como VIAF. Las propiedades del modelo al efecto son `authorityAssigner`, que indica la entidad que asigna la materia y `authoritySource`, que indica la lista origen de esa materia.

Los metadatos administrativos también tienen cabida en el modelo a través de información de carácter general como `creationDate`, `changeDate`, `descriptionConventions`, `descriptionLanguage`, `descriptionSource`, etc.



Figura 13 Modelo de datos de BIBFRAME. Fuente: ISQ NISO 2013.

3.2.8.3 Representación de obra en BIBFRAME

Los IRIs propuestos para Biblioteca UPM son ficticios y se utilizan a modo de ejemplo.

```
@prefix bf: <http://bibframe.org/vocab/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
```

#Obra#

```
< http://upm.es/library/cat/item/6000084373> a bf:Work,
    bf:Text;
    bf:workTitle [ a bf>Title ;
        bf:titleValue "Tecnologías de la web semántica" ];
    bf:authorizedAccessPoint "Tecnologías de la web semántica" ;
    bf:classificationUdc [ a bf:Classification ;
        bf:classificationNumber "004.738.52" ;
        bf:classificationScheme "classificationUdc" ] .
    bf:creator [ a bf:Person ;
        bf:authorizedAccessPoint "Pastor Sánchez, Juan Antonio";
        bf:hasAuthority [ a madsrdf:Authority ;
            madsrdf:authoritativeLabel "Pastor Sánchez, Juan Antonio" ;
            bf:label "Pastor Sánchez, Juan Antonio" ] ;
        bf:derivedFrom <http://upm.es/library/cat/item/6000084373> ;
        bf:hasAnnotation [ a bf:Summary ;
            bf:annotates <http://upm.es/library/cat/item/6000084373> ;
            bf:label "Descripción del editor" ;
```

#Direcciones ficticias de sumarios, tablas de contenido y ficheros de texto en el catálogo#

```
    bf:review < http://upm.es/library/cat/summary/item/6000084373> ],
    [ a bf:TableOfContents ;
        bf:annotates <http://upm.es/library/cat/item/6000084373> ;
        bf:label "TOC" ;
        bf:tableOfContents <http://upm.es/library/cat/toc/item/6000084373> ],
    [ a bf:Annotation ;
        bf:annotates <http://upm.es/library/cat/item/6000084373>;
        bf:annotationBody <http://upm.es/library/cat/textfile/item/6000084373>;
        bf:label "Sample text" ] ;
```

#Instancia#

```
    bf:hasInstance [ a bf:Instance,
        bf:Monograph ;
```

#Volumen#

```
    bf:carrierCategory < http://rdvocab.info/Elements/carrierType/nc> ;
    bf:derivedFrom < http://upm.es/library/cat/item/6000084373> ;
    bf:dimensions "18 cm. " ;
    bf:extent "120 p." ;
    bf:heldBy [ a bf:heldBy ;
        bf:label "004.738.52" ;
        bf:shelfMarkUdc "004.738.52" ] ;
    bf:instanceOf < http://upm.es/library/cat/item/6000084373> ;
    bf:instanceTitle [ a bf>Title ;
```

```

        bf:titleValue "Web semántica" ] ;
        bf:isbn13 <http://isbn.example.org/ 9788497884747> ;
#Medio textual#
        bf:mediaCategory < http://rdvocab.info/Elements/mediaType/n> ;
        bf:note "Aparece un enlace para descarga gratuita de un ebook en format
        epub p. 118" ;
        bf:providerStatement "Barcelona, Editorial Universidad Oberta de Cataluña,
        El profesional de la Información, ©2011." ;
        bf:publication [ a bf:Provider ;
                bf:providerDate "@2011." ;
                bf:providerName [ a bf:Organization ;
                        bf:label "Editorial UOC" ] ;
                bf:providerPlace [ a bf:Place ;
                        bf:label "Barcelona" ] ] ;
        bf:responsibilityStatement "Pastor Sánchez, Juan Antonio" ;
        bf:language "@eng" ;
        bf:subject [ a bf:Topic ;
                bf:authorizedAccessPoint "Web semántica" ;
                bf:hasAuthority [ a madsrdf:Authority,
                        madsrdf:ComplexSubject ;
                        madsrdf:authoritativeLabel "Web semántica"@sp ] ;
                bf:label "Web semántica" ] ;
        bf:title "Tecnologías de la web semántica" .

```

3.3 CASOS DE USO

3.3.1 LIBRARY LINKED DATA INCUBATOR GROUP REPORT: USES CASES

Otra de las actividades del Grupo del W3C en sus estudios sobre Library Linked Data ha sido la investigación sobre casos significativos y considerados como ejemplos de buenas prácticas respecto al uso de tecnologías Linked Data en bibliotecas. El grupo ha establecido ocho áreas de análisis y ha presentado conclusiones y aspectos de mejora de gran interés respecto a las actividades que se desarrollan dentro de ellos (Vila-Suero, 2011):

Aplicación de Linked Data a los registros bibliográficos:

1. Estandarización de los registros bibliográficos.
2. Mantenimiento de registros más eficaz y menos costoso.
3. Mejora en los sistemas de recuperación de la información.
4. Enriquecimiento de los registros por fuentes externas vinculadas.

Aplicación de Linked Data a los datos de autoridades:

1. Interfaces interactivas basadas en ficheros de autoridad que mejoran la precisión de la búsqueda al permitir el acceso a cualquier metadato del sistema y promoviendo el desarrollo del multilingüismo en los ficheros de autoridades.
2. Agregación de diferentes registros de autoridad con contenidos complementarios aptos para enriquecer el proceso descriptivo.

Aplicación de Linked Data a los vocabularios controlados:

1. Mejoras en el enriquecimiento y la recuperabilidad de la información, permitiendo búsquedas temáticas, recuperación de resultados en distintos idiomas, navegación transversal por diferentes disciplinas enlazadas, etc.
2. Reutilización de vocabularios mediante la extensión de vocabularios centrales a otros más específicos.
3. Servicios de alineamiento de vocabularios para la gestión de actualizaciones que ofrecen versiones estables prácticamente en tiempo real.

Aplicación de Linked Data a los archivos documentales:

1. Conexión de diferentes archivos a través de la vinculación de sus metadatos: Interoperabilidad archivística.
2. Mejoras en la recuperabilidad de los documentos.
3. Mejoras en la gestión de datos, preservación e interoperabilidad de datos heterogéneos, muy comunes en los fondos de archivo.

Aplicación de Linked Data a las referencias y citas documentales:

1. Mejoras en la información que ofrecen las publicaciones en sus referencias y citas, accediendo directamente a la información referenciada y perfeccionando la navegabilidad entre documentos citados (*Enhanced publications*).
2. Aplicaciones a la valoración según citas, aportando nuevos índices de impacto.
3. Recuperación de publicaciones del contexto de la cita mediante tecnologías semánticas.

Aplicación de Linked Data a los objetos digitales:

1. Los datos vinculados permiten la agregación de recursos digitales por los usuarios consumidores de la información.
2. Enriquecimiento a través de la vinculación a otros datos complementarios.
3. Posibilidad de reutilización de los materiales digitales vinculados o sus metadatos.

Aplicación de Linked Data a las colecciones de las bibliotecas:

1. Descripciones semánticas a nivel de colección que permiten la generación de nuevos metadatos globales lo que supone la identificación de la colección como un objeto unitario.
2. Clasificación de colecciones por usuarios finales.
3. Localización de colecciones y aplicaciones móviles sobre las mismas.

Aplicación de Linked Data a los usos sociales en el contexto bibliotecario y a utilizaciones innovadoras de la información:

1. Agregación de la información derivada de los usos sociales en la biblioteca: información de usuarios o aportada por ellos, información de uso de los recursos de la biblioteca.
2. Publicación y agregación distribuida de la información social.
3. Los usos innovadores suelen requerir datos legibles por máquinas, cuestión facilitada por las tecnologías de datos vinculados.

3.3.2 EUROPEANA

Europeana es el proyecto más ambicioso de difusión del patrimonio cultural europeo. Actúa sobre la base de todo tipo de recursos culturales de todos los países de la Unión Europea. Ofrece un modelo de datos estructurado bajo Linked Data, que permite la reutilización de metadatos de los proveedores de recursos culturales, estableciendo un marco de interoperabilidad de esos recursos bajo la filosofía Open Data en el ámbito cultural y europeo. La intercomunicación entre los distintos agentes del modelo ha sido un objetivo de difícil consecución, dada la disparidad de proveedores de recursos culturales, lo que ha supuesto procesar una gran variedad de metadatos de muy diferentes orígenes, con diferente nivel de granularidad en la descripción y las múltiples particularidades respecto a la propiedad intelectual de los recursos suministrados.

El *Europeana Data Model (EDM)* está desarrollado mediante la familia de lenguajes RDF y presenta las siguientes características generales (Isaac, Clayphan, & HasLhofer, 2012):

1. Distinción precisa entre elemento real y su representación digital.
2. Separación entre el elemento y sus metadatos descriptivos.
3. Permite incluir varias declaraciones sobre un mismo recurso, aunque sean contradictorias.
4. Proporciona también soporte a los recursos contextuales como por ejemplo elementos de vocabularios controlados.
5. Posee un modo de representación de objetos abierto, flexible y cuenta con la posibilidad de descripción a diferentes niveles, lo que permite el tratamiento de objetos digitales complejos.

6. Preferencia por estándares establecidos y vocabularios de referencia: OAI-ORE (Open Archives Initiative Object Reuse & Exchange), Dublin Core, SKOS y CIDOC-CRM.
7. Generación de entidades EDM como objetos y agregaciones de datos y asignación de IRIs a dichas entidades para su identificación.
8. Enriquecimiento semántico (por parte de Europeana) de los metadatos de proveedores de recursos culturales a través de colecciones de elementos semánticos como: GeoNames (ubicaciones), Gemet (temas genéricos), Semiun Time Ontology (valores temporales) y Dbpedia (por ahora sólo para valores de personas).
9. Los IRIs incluidos en los metadatos son utilizados para crear enlaces con otros servicios Linked Data que contengan información agregada al objeto descrito, ya desde otros proveedores, ya desde editores de valores reconocidos.
10. Incluye información de procedencia a través de los metadatos Provenance.

3.3.2.1 Representación sintáctica del modelo de datos de Europeana (EDM)

Como se ha referido, EDM permite establecer diferentes niveles de granularidad en la descripción de recursos culturales, pudiendo además especificar perfiles de aplicación específicos según el contexto de los objetos. La aplicación de Linked Data aporta al modelo beneficios en la calidad de la recuperación de sus recursos, enriquecimiento automático de los mismos y aportación a la nube de datos de metadatos y recursos de gran calidad. La distribución de sus clases principales se muestra en el siguiente gráfico:

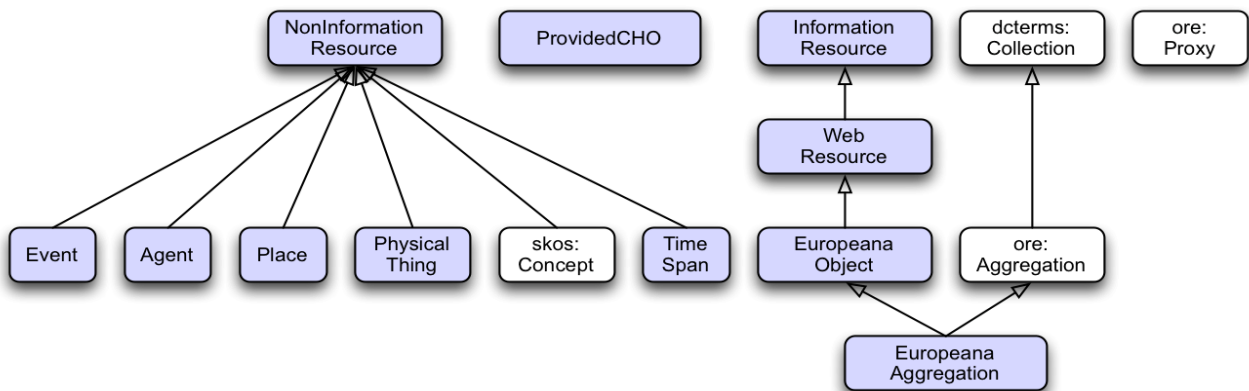


Figura 14 Esquema de clases del modelo de datos de Europeana. Fuente: Europeana (2013).

El gráfico siguiente presenta las principales propiedades del modelo de datos de Europeana:

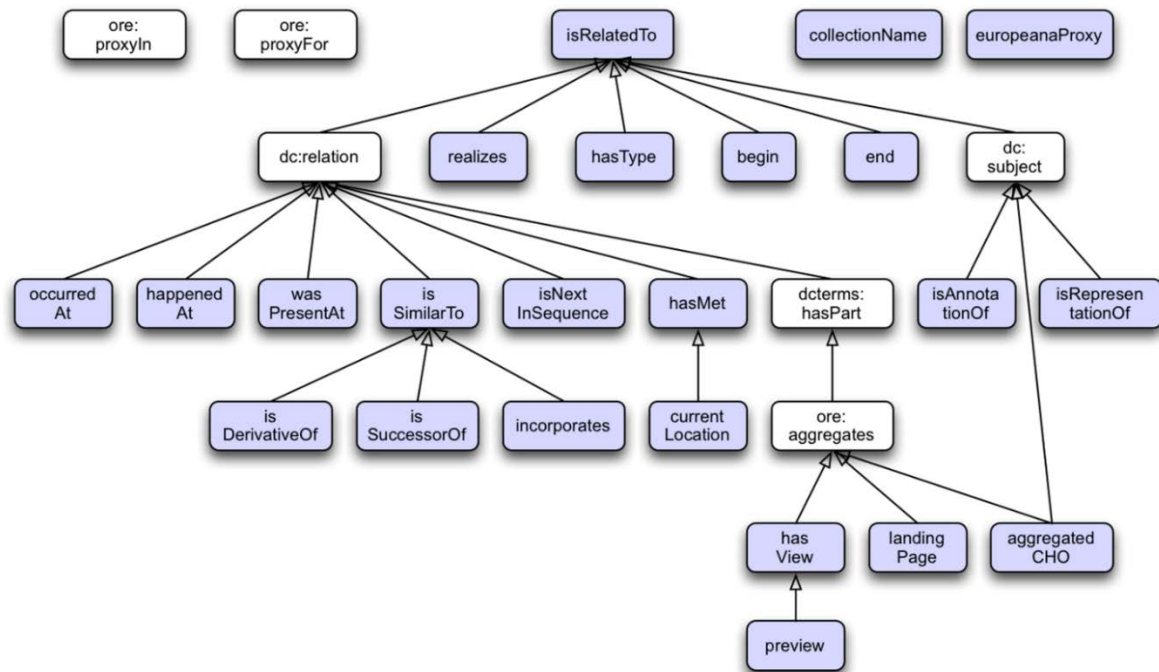


Figura 15 Esquema de propiedades del modelo de datos de Europeana. Fuente: Europeana (2013).

Los tres elementos fundamentales del EDM son (Isaac et al., 2012; Isaac et al., 2013):

1. Metadatos descriptivos de un objeto del patrimonio cultural que el proveedor entrega a Europeana: `edm:ProvidedCHO`.
2. Representaciones de los objetos digitales mediante recursos web: `edm:WebResource`.
3. Agregaciones o representación del conjunto de recursos referidos a un objeto digital suministrado por el proveedor, incluyendo los metadatos descriptivos de todo el conjunto de recursos: `ore:Aggregation`

EDM propone tres aproximaciones de descripción: un enfoque descriptivo orientado al objeto, el enriquecimiento de datos con clases conceptuales y un último enfoque orientado a eventos. La descripción enfocada a objetos se realiza mediante representaciones ligadas directamente a el objeto y fundamentalmente expresadas con Dublin Core (`dc:` y `dcterms:`). El segundo enfoque es el que asocia a los datos las entidades contextuales:

1. `edm:Agent`: para la representación de personas o entidades colectivas.
2. `edm:Place`: para entidades relativas a datos geográficos.
3. `edm:TimeSpan`: para períodos de tiempo o fechas.
4. `skos:Concept`: para la estructuración de KOS, como tesauros, taxonomías, etc.

El tercer enfoque describe los objetos mediante la caracterización de los eventos en los que el objeto ha participado. Estos eventos relacionan el objeto con otras entidades mediante la clase `edm:Event`. Las propiedades que representan estas relaciones son:

1. `edm:wasPresentAt`: indica la relación entre el objeto digital y el evento.
2. `edm:happenedAt`: indica la relación evento-lugar.
3. `edm:occurredAt`: indica la relación que se establece entre el evento y el plazo temporal en el que sucedió.

Europeana Data Model propone un esfuerzo mayor de descripción a los proveedores propiciando que sus contribuciones al haber de Europeana sean más semánticas utilizando el esquema RDFS, lo que permite ofrecer jerarquías estructuradas de metadatos que describan más adecuadamente el contexto de los recursos culturales a través de la definición de niveles de clases y propiedades (HasLhofer & Isaac, 2014; National Library of the Netherlands & European Commission, 2014).

```
<rdf:Property
rdf:about="http://cdp.upm.es/R/?object_id=479236"/features>
<rdfs:subPropertyOf
rdf:resource="http://purl.org/dc/elements/1.1/description"/>
</rdf:Property>
```

En el ejemplo anterior se establece una jerarquización de los elementos descriptivos, se declara el recurso (una brújula taquimétrica excéntrica de la Colección Digital Politécnica) que contiene las características del objeto real identificado por un IRI y esta declaración se subordina mediante la propiedad `"rdfs:subPropertyOf"` a la propiedad `"description"` de Dublin Core, permitiendo que las características del aparato puedan ser recuperadas en el contexto de la descripción general.

Otra de las posibilidades de enriquecer los objetos con conjuntos de metadatos desde fuentes diferentes es el *Proxy*. Cabe la posibilidad de que dos proveedores envíen cada uno su propia y diferente representación digital sobre un mismo recurso, cada una de ellas supone una agregación. Estas diferentes agregaciones dan lugar a un *Proxy* (expresado mediante `ore:proxy`) diferente para cada una de ellas. El *Proxy* define el objeto provisto desde el punto de vista del proveedor del recurso. Esta estrategia es importante para el concepto de Europeana, pues permite que convivan descripciones parciales y potencialmente diferentes de un mismo recurso, lo que evita la eliminación de un proveedor y por tanto se conserva la importante información de procedencia y la riqueza que aportan ambas descripciones.

El proxy se vincula al objeto mediante `ore:proxyFor` y a la agregación del proveedor mediante `ore:proxyIn`. Si existen varios *Proxy*, todos ellos se conectan al recurso `edm:ProvidedCHO` cuya representación es ajena a los diferentes puntos de vista de los sucesivos proveedores (al estilo de la "obra" FRBR).

A la par que el enriquecimiento externo ya comentado, Europeana ofrece un enriquecimiento interno a través de `edm:EuropeanaAggregation`, que permite la inclusión de información sobre propiedad intelectual, restricciones de acceso o información relevante de interés.

La estructuración de los metadatos en jerarquías es otra de las facetas avanzadas de EDM. Ya hemos referido arriba como la introducción de clases y subclases o propiedades y subpropiedades mejoraban su semántica. EDM también cuenta con elementos propios que permiten definir jerarquía de datos apoyándose en vocabularios externos: `dcterms:hasPart` y `dcterms:isPartOf` o `edm:isNextInSequence` para introducir una ordenación de las partes del objeto en su caso. Otra posibilidad de establecer relaciones enriquecedoras se define por `edm:isRelatedTo` que permite establecer vínculos entre dos objetos que representan el mismo ente lógico (Isaac et al., 2013; National Library of the Netherlands & European Commission, 2014).

3.3.2.2 Descripción completa de un recurso de la Colección Digital Politécnica

```
<rdf:RDF xsi:schemaLocation="http://www.w3.org/1999/02/22-rdf-syntax-ns# http://www.europeana.eu/schemas/edm/EDM.xsd">
```

```
<ore:Aggregation
rdf:about=http://cdp.upm.es/R/?object_id=479236&func=dbin-jump-
full#aggregation/>
<edm:aggregatedCHO rdf:resource="
http://cdp.upm.es/R/?object_id=479236"/>
<ore:aggregates>
<edm:hasView>
<edm:WebResource
rdf:about="http://cdp.upm.es/webclient/DeliveryManager?pid=479236&custo
m_att_2=simple_viewer/>
</edm:hasView>
</ore:aggregates>
<edm:dataProvider>BIBUPMMAD</edm:dataProvider>
<edm:isShownAt
rdf:resource=" http://cdp.upm.es/R/?object_id=479236"/>
<edm:provider>Biblioteca Universitaria Campus Sur UPM</edm:provider>
<edm:rights rdf:resource=" http://creativecommons.org/licenses/by-nc-
nd/3.0/es/" />
</ore:Aggregation>

<edm:ProvidedCHO rdf:about="http://cdp.upm.es/R/?object_id=479236"/>
<edm:hasMet>
<dcterms:temporal>Siglo XX</dcterms:temporal>
<dc:language>spa</dc:language>
```

```

<dc:date>2011</dc:date>
<dc:creator>Biblioteca Universitaria Campus Sur</dc:creator>
<dc:creator>Avila-R</dc:creator>
<dcterms:provenance>ETSI Topografía, Geodesia y
Cartografía</dcterms:provenance>
<dcterms:provenance>Museo de Instrumentos
Topográficos</dcterms:provenance>
</edm:hasMet>
<dcterms:description>Brújula taquimétrica / taquímetro repetidor de
anteojo excéntrico, de doble lectura directa en nonios para el limbo
descubierto H y simple en el limbo descubierto V y gran brújula
central. Anteojo con nivel tubular para función de equialtímetro.
Nivelación con nivel esférico. La descripción completa del aparato se
puede consultar en el archivo adjunto: "Ficha técnica brújula
Breithaupt". Sin caja. <dcterms:description>
<ex:features>incorpora un retículo con filos estadimétricos y limbos
acimutales de vidrio</ex:features>
<dc:title>Brújula taquimétrica excéntrica Breithaupt & Sohn</dc:title>
<dcterms:medium>recurso en línea</dcterms:medium>
<dc:subject
rdf:resource="http://www.upm.es/biblioteca/kos/sh/brújulas"/>
<dc:subject rdf:resource="http://www.upm.es/biblioteca/kos/sh/aparatos
topográficos"/>
<edm:hasType>
<dc:type>taquímetro repetidor</dc:type>
</edm:hasType>
<dc:rights>Reconocimiento - NoComercial - SinObraDerivada (by-nc-
nd)</dc:rights>
<edm:type>Physical Object</edm:type>
<edm:wasPresentAt>
<edm:event>1ª Exposición de material patrimonial topográfico de la
UPM</edm:event>
</edm:wasPresentAt>
</edm:ProvidedCHO>
</rdf:RDF>

```

3.3.3 BIBLIOTECA NACIONAL DE ESPAÑA. DATOS BNE 2.0

La Biblioteca Nacional de España en colaboración y el Ontology Engineering Group (OEG) de la Universidad Politécnica de Madrid, llevan años colaborando para el desarrollo de productos bibliotecarios publicados en Linked Data. El proyecto, en constante evolución, tiene como objetivo publicar en Linked Data los datos del catálogo BNE, transformando los registros MARC en tripletas RDF mediante mapeos automáticos de los datos. El proyecto se manifiesta como un

ejemplo prototípico de colaboración y coordinación transversal de diferentes grupos profesionales: desde bibliotecarios de diversas especialidades hasta los ingenieros que aportaban sus conocimientos sobre sistemas, demostrando que el ecosistema Linked Open Library Data debe construirse desde la conjunción de conocimientos y competencias.

El primer grupo de registro se publicó en el año 2011. Como requisito necesario del proyecto se planteó la intención de utilizar estándares IFLA como FRBR e ISBD. El OEG estableció una ruta de desarrollo que abarcaba:

1. La especificación de los datos bibliográficos.
2. Modelado en FRBR.
3. Generación de tripletas RDF mediante el software de desarrollo propio “MARIMBA”.
4. Establecimiento de vínculos con otros datasets como la Biblioteca Nacional Alemana, el catálogo sueco LIBRIS, el directorio de autoridades VIAF, etc.
5. Publicación de los datos vinculados mediante el registro en CKAN, generación de un *sitemap* y su publicación en Google y en Sindice.
6. Creación de un entorno de explotación de datos mediante la creación de un site y el establecimiento de un SPARQL endpoint.

En abril de este mismo año, la BNE ha presentado la evolución de su proyecto. Sus principales novedades son las siguientes:

1. Creación de la ontología BNE (Biblioteca Nacional de España, 2014a), que continua y aumenta la utilización de estándares bibliotecarios como RDA, SKOS, DC, MADS y BIBO. La extensión a los estándares se efectúa mediante `owl:subClassOf` y `owl:subPropertyOf`. También la actualización del proyecto introduce de modo más consistente el multilingüismo e incorpora datos de procedencia (Provenance). La ontología BNE se declara como un sistema estable y usable, estando documentada en profundidad. Su modelo básico de datos responde al siguiente gráfico.
2. Ampliación de los registros MARC procesados, inclusión de las etiquetas de idioma en los títulos “@es”, miniaturas en los recursos bibliográficos e inclusión de datos como el Depósito legal o el tipo de material (RDA).
3. Ampliación de los registros de autoridad.
4. Aumento de la mapeos con datasets externos: Geonames, Library of Congress, ISNI.
5. Enriquecimiento de datos con la conexión a la Biblioteca Digital Hispánica y Dbpedia.
6. Integración de datos de biografías y miniaturas de autores.
7. Inclusión de metadatos de procedencia (Provenance) y OAI-ORE (Open Archives Initiative Object Reuse and Exchange).
8. Adaptación del portal a diferentes dispositivos de representación.
9. Los datos están publicados de acuerdo a las especificaciones y condiciones de la licencia Creative Commons Zero (CC0).

El proyecto Datos BNE 2.0 no se puede dar como finalizado, su estado es de permanente actualización tanto de técnicas y procesos como de contenidos. En breves fechas se inaugurará el nuevo portal de datos semánticos de la BNE, cuya interfaz permite disfrutar de una experiencia innovadora en la recuperación de la información. (Biblioteca Nacional de España, 2014b; Vila-Suero, Villazón-Terrazas, & Gómez-Pérez, 2012; Vila-Suero, 2014) .

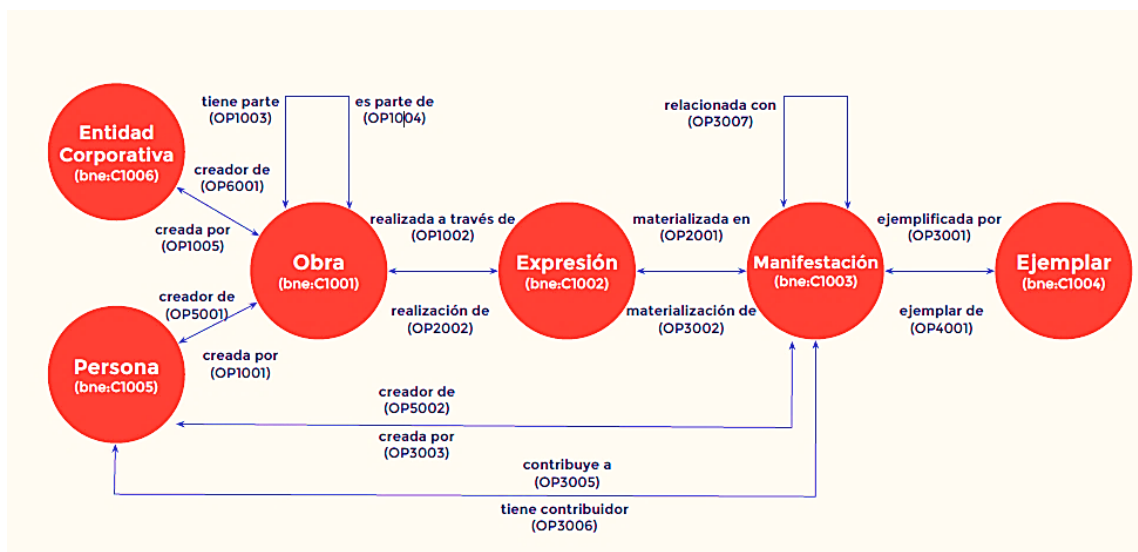


Figura 16 Ontología de la Biblioteca Nacional de España. Fuente: Ontology Engineering Group (2014).

3.4 BÚSQUEDA SEMÁNTICA DE LA INFORMACIÓN. HERRAMIENTAS DE DESCUBRIMIENTO

Las tecnologías semánticas también afectan a los sistemas de recuperación de la información. Por ahora, los catálogos son accesibles vía Web pero únicamente para consultar la base de datos local, es decir que los datos permanecen en silos y sólo circulan en el ámbito de la propia biblioteca o a lo sumo de un consorcio de ellas. Linked Data pretende llevar la información que contienen las bibliotecas a la Web de datos y ofrecer un producto de salida que permita su explotación por cualquier agente que lo necesite. Visto desde otro punto de vista, la biblioteca puede aprovechar los datos externos y enriquecer con ellos la información suministrada por los catálogos.

El OPAC representa el modelo tradicional de búsqueda de información en la biblioteca. Su versión electrónica, nacida no hace muchos años, está pasando por la misma época de transformación que el resto de las estructuras y sistemas bibliotecarios, su misión como recuperador de la información del sistema local está pendiente también de un proceso renovador.

Desde un punto de vista histórico, el catálogo no ha dejado nunca de evolucionar: originalmente ha permitido la búsqueda y recuperación de la información en los fondos físicos de la biblioteca, y con el transcurso del tiempo y la evolución de las tecnologías de la información, ha ido aumentando su ámbito de recuperabilidad, desde los recursos electrónicos almacenados en bases de datos on line, hasta su integración con otros sistemas de descubrimiento de la biblioteca, como las herramientas de metabúsqueda, búsqueda federada, agregadores de información, información en dispositivos móviles, servicios de disseminación selectiva de la información, etc. (Breeding, 2014; Iacono, 2013; Le Boeuf, 2013; Picco & Ortiz-Repiso, 2012a).

La llegada al escenario bibliotecario de las FRBR ha provocado un replanteamiento general aunque muy lento de las funciones del catálogo. FRBR tiene en cuenta las tareas del usuario para la recuperación de la información, lo que ha supuesto un soporte para la adaptación de los OPAC,s de nueva generación a la hora de incorporarse al modelo bibliográfico conceptual, ofreciendo nuevos modos de representar la información, nuevas herramientas similares a los motores web, servicios que permiten la gestión de la información por el usuario y otros recursos enriquecedores de la información (Iacono, 2013; Picco & Ortiz-Repiso, 2012a).

El proceso de búsqueda tradicional cuenta con importantes obstáculos que dificultan el éxito en la recuperación de la información, desde los que interpone el propio sistema, hasta las propias características del usuario que busca la información (Breeding, 2014). Las ventajas que a priori ofrece un sistema de recuperación de la información de recursos publicados en Linked Data apuntan hacia una estructura de más fácil navegación, con un estilo de búsqueda más orientado al entorno Web, permitiendo al usuario un aprendizaje de las técnicas de recuperación en el nuevo contexto, minorando las barreras cognitivas implicadas en los procesos de búsqueda y mayores posibilidades de extracción de la información en búsquedas semánticas, dada la nueva naturaleza enriquecida de la información que ofrecen los datos vinculados. El usuario en sus ecuaciones de búsqueda puede beneficiarse de la superioridad de indización de los vocabularios controlados y las ontologías, (Iacono, 2014); por ejemplo, la utilización de encabezamientos de materia en Linked Data, permite descubrir nuevas fuentes de información, dada la rica estructura de relaciones entre conceptos, pues a las relaciones internas del vocabulario se suman las externas establecidas mediante mapeos, los cuales permiten interconectar diferentes sistemas de indización incluso en diferentes idiomas.

Las conexiones múltiples entre los temas buscados o autores implicados en una investigación pueden llegar a revelar información de más calidad o pertinencia, ayudando a verificar la repercusión de un cierto trabajo o creador. Hay que recordar que Linked Data procesa los datos también para la exploración automática por agentes semánticos, así, el empleo de herramientas de descubrimiento automáticas nos proporciona nuevas posibilidades de capturar la información de modo directo o inferido (Iacono, 2014).

La gestión de las materias en las bibliotecas y por ende en los catálogos adoptará una nueva dimensión: materias locales expuestas como Linked Data pueden establecer relaciones con otras materias en otros sistemas y generar archivos de autoridad que al aplicarse al registro de

catalogación ofrezcan una multitud de puntos de acceso y por ello de posibilidades de ser recuperados en una búsqueda del catálogo. La búsqueda por materias puede dejar de ser la menos satisfactoria para el usuario y más si se conectan y alinean vocabularios controlados como los referidos, con otros como sistemas de términos provenientes de la clasificación por los propios usuarios como las folksonomías. Finalmente la recuperabilidad de la información puede mejorarse a través de conceptos relacionados con aquellos que primariamente hemos buscado, es la denominada “red de significados” que permite extraer la información de diversas comunidades cuyo conocimiento es complementario pero que se genera desde diferentes puntos de vista (Iacono, 2014).

Otros sistemas de descubrimiento ofrecen links de interés mediante listas de recomendación producidas a través de metadatos vinculados y ontologías. Estos sistemas de enriquecimiento semántico de las búsquedas se denominan *recommendation systems* y ofrecen al usuario información adicional de su interés. Las ontologías completan estos sistemas de información representando preferencias de usuarios y perfiles de búsqueda (Hyvönen, 2012b).

En un contexto como el descrito, las posibilidades futuras no son previsibles: la interconexión de todos los datos nos llevaría al enriquecimiento de datos bibliográficos con datos museísticos, datos de geolocalización, biográficos etc. El ámbito de los recursos crece con cada set de datos publicados en la nube de la Web semántica, los catálogos pueden generar dinámicamente recursos de interés partiendo de una búsqueda sobre un tema concreto, es el caso de la Biblioteca Nacional Francesa que genera tras la visualización de una entidad una serie de enlaces a recursos muy variados como películas, imágenes etc., sobre dicha entidad. Además es posible que el catálogo deje de ser una herramienta reconocida de la biblioteca, pues en realidad ejecutará su función en la web y el usuario que lo utilice, quizás pierda la noción de que está ante una herramienta de la biblioteca (Bermés, 2013; Iacono, 2014; Le Boeuf, 2013).

Veamos con más detalle el caso anteriormente mencionado de la BNF. La Biblioteca Nacional de Francia ha desarrollado un proyecto de catálogo que soportar estructuras semánticas de datos vinculados. La propia biblioteca define los objetivos que ha pretendido conseguir con su desarrollo:

1. Facilitar las búsquedas desplegando resultados en torno a la expresión “Obra” de FRBR.
2. Utilizar técnicas de datos vinculados para combinar documentos de varias fuentes en una misma lista.
3. Proponer vínculos con los recursos disponibles en abierto.
4. Enriquecer la lista de recomendaciones desde repositorios Open Data.
5. Proponer la búsqueda por materias a través de la representación gráfica de las materias de RAMEAU.
6. Utilizar las mismas técnicas y servicios que las webs comerciales.

Este proyecto contempla la implantación del catálogo general Open Cat en las bibliotecas locales, con acceso Linked Data a los fondos de la BNF y completada y enriquecida la información con recursos alojados localmente. Open Cat es compatible con los sistemas locales basados en MARC (Le Boeuf, 2013).

Para completar un panorama real de la situación se deben evaluar los obstáculos que las tecnologías semánticas están encontrándose en el camino hacia un OPAC semántico (Iacono, 2014):

1. La escasa penetración del modelo FRBR en los catálogos en general.
2. El modelo de gestión de datos en la búsqueda se mantiene en la esfera “interna” del ámbito bibliotecario, no hay un movimiento reconocible hacia la extensión de los sistemas de búsqueda en contextos de datos vinculados.
3. La clonación de los sistemas de búsqueda web no está todavía bien definida.
4. La falta de competencias necesarias para trabajar en el catálogo en modo semántico.
5. Recursos económicos y personales escasos.
6. Escasa percepción general de lo que las tecnologías semánticas pueden aportar a la búsqueda de información.

Finalmente hacer referencia a algunas soluciones de búsqueda semántica disponibles actualmente en el mercado. Destaca la herramienta de búsqueda VuFind, (Villanova University's Falvey Memorial Library, 2014), un software Open Source que permite la búsqueda en todos los recursos de la biblioteca (externos e internos), fácilmente integrable con la interfaz del catálogo en línea. Utiliza lenguajes de descripción semántica como Schema.org para la recuperación. VIVO (Corson-Rikert & Hidalgo, 2014) es otra novedosa herramienta, que ofrece una plataforma semántica y Open Access que permite la recuperación de información y de los resultados de la investigación en múltiples disciplinas con soporte semántico (Breeding, 2014).

Una de las principales plataformas de búsqueda de datos vinculados es, sin lugar a dudas Síndice (DERI, Fondazione Bruno Kessler, & OpenLink software, 2014), proyecto soportado por DERI y OPENLINK, entre otros y que ofrece, además de búsqueda en cualquier formato semántico distribuido por la Web, API,S para implementar sistemas de búsqueda semántica y herramientas de descubrimiento para datos vinculados (Méndez, 2010). La recuperabilidad de los set de datos no es un procedimiento tan perfeccionado como el de las IRIs desreferenciables; es por ello que se requieren métodos adicionales para indizar y rastrear los set de datos, al menos hasta que se generalice la asignación de metadatos descriptivos a los conjuntos de datos vinculados. Síndice lleva a cabo estas funciones, indiza y rastrea, y además, proporciona una API que permite a los desarrolladores detectar datos relevantes y utilizarlos en sus aplicaciones. Síndice utiliza tres modos de indización: acumula documentos RDF y crea un índice sobre ellos identificado por un IRI, también utiliza asignación de descriptores a los set de datos y técnicas IFP (Inverse Functional Properties), que son objetos de tripletas RDF que se vinculan única y exclusivamente a un solo

sujeto de la tripleta, especificándolo de modo unívoco) (Oren, Delbru, Catasta, Cyganiak, & Tummarello, 2008).

Una breve pero especial referencia se debe hacer a cerca de LIBRIS, el catálogo de la Swedish University (Swedish University, 2014a) y las bibliotecas de investigación suecas, gestionado por la “National Library of Sweden”. El proyecto de catálogo semántico ha prestado especial atención a las tácticas de búsqueda de los usuarios, ha testado el producto con ellos y ha realizado entrevistas para obtener información de primera mano, todo ello con la pretensión de proveer una experiencia de búsqueda más eficiente. Para fomentar la interoperabilidad tanto respecto a protocolos veteranos como a los nuevos estándares, los desarrolladores han implementado una amplia variedad de protocolos desde Z39.50 hasta los propios de Linked Data (Bermés, 2013).

Su diseño presenta múltiples características avanzadas: importación y exportación de datos automatizada, integración con API,s, especialmente las disponibles en formatos abiertos, utilización de identificadores específicos para los recursos bibliográficos (LIBRIS-ID), autoridades estructuradas bajo FOAF y SKOS, registros bibliográficos bajo BIBO y elementos DC, mapeos con Dbpedia, LCSH subject headings en Linked Data VIAF, Open Library, etc.

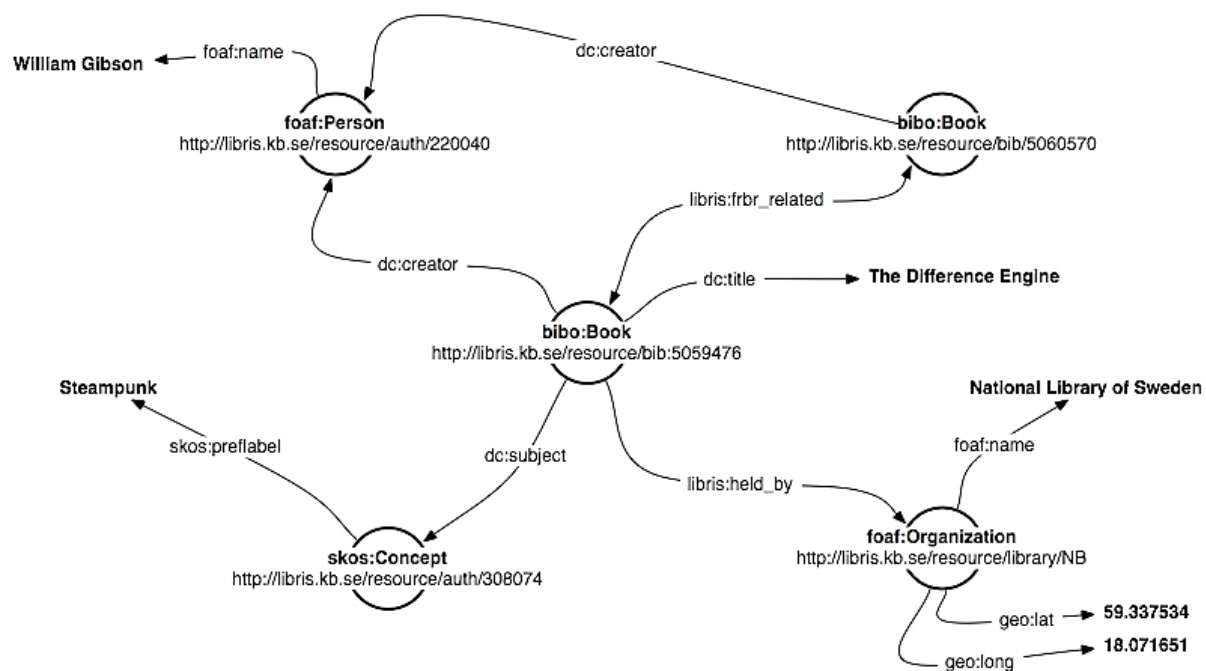


Figura 17 Modelo de datos del catálogo LIBRIS. Fuente: Swedish University (2014)

Finalmente indicar que el W3C hace un seguimiento de las herramientas de descubrimiento centradas en datos y con funcionalidades semánticas en su página web (W3C, 2014).

3.5 UN ANÁLISIS CRÍTICO DE LAS TECNOLOGÍAS LINKED DATA EN BIBLIOTECAS

Apuntadas las mejoras y posibilidades que las tecnologías semánticas nos ofrecen, cabe ahora abordar algunos de los puntos oscuros cuya presencia distorsiona las expectativas de futuro de Linked Data en el contexto LODGLAM.

Clayton M. Christensen en 1997 definió la teoría sobre la innovación disruptiva, teoría que hace referencia al papel que algunas tecnologías emergentes desempeñan en un dominio determinado, ocupando el puesto de tecnologías más obsoletas. Para Christensen, Linked Data no es una tecnología disruptiva, y son varias las debilidades que presenta: por un lado no es una tecnología madura, tampoco puede, a día de hoy, ser sustituta de modo completo, de las tecnologías que se utilizan en bibliotecas; del mismo modo, no termina de introducirse en los canales comerciales ni en general en los servicios de información (Moulaison & Million, 2014).

Todo este cúmulo de inconvenientes debe solucionarse para propiciar la impulsión definitiva de las tecnologías semánticas. Un trabajo más coordinado a nivel internacional, donde lo público y lo privado converjan, puede ser la solución (Moulaison & Million, 2014; Saorín, Peset, & Ferrer-Sapena, 2013; Van-Hooland, Verborgh, & Van-de-Valle, 2012). Los puntos débiles que se manifiestan y que deben ser corregidos a corto plazo son los siguientes:

1. Respecto a los identificadores unívocos (IRIs), parece alarmante la gran cantidad de enlaces rotos incluso en servicios contrastados como The DataHub. La proliferación de IRIs diferentes para los mismos recursos también aparece con demasiada frecuencia. Este es un problema fundamental sobre todo en la aplicación LOD a bibliotecas.
2. En cuanto a los estándares, no se debe considerar RDF o SPARQL como estándares obligatorios, otros son posibles candidatos y este tema está todavía abierto. A veces las propias organizaciones se dejan llevar por el entusiasmo de los datos vinculados, así por ejemplo, en la web de Schema.org no se hace referencia a que su estándar sea considerado como Linked Data, mientras que la reciente catalogación de millones de recursos por OCLC con Schema.org, ha sido definida y proclamada como la mayor publicación de registros bibliográficos en Linked Data por la comunidad semántica.
3. La recuperación de la información desde datos estructurados en Linked Data está lejos de ofrecer valores similares a la Web.
4. El almacenamiento de datos en *triplestores* tiene algunos inconvenientes que resolver: no parece fácil encontrar cómo están estructurados y dónde están, aunque se anuncie todo lo contrario.
5. Es cierto que, en las investigaciones en datos vinculados, normalmente basadas en prototipos y desplegadas sobre volúmenes de datos no muy amplios, los resultados de las tecnologías de vinculación de datos son esperanzadores. En el mundo real, en cambio, no hay establecido un sistema de control estándar de calidad de la web y la gestión de

datos y su preservación están aún en pañales. Todo ello puede redundar en pérdidas de calidad en las publicaciones, desde el mismo momento en que los datos vinculados se encuentren con los mismos problemas que el resto de participantes en la Web. Además LD se manifiesta como un conjunto de técnicas complejas, lo que puede ser un impedimento muy relevante para su definitiva implantación, ya que la sencillez ha sido la nota común en los estándares que han triunfado en la Web.

6. Existe cierta desconfianza en las bibliotecas sobre sus datos expuestos en almacenes externos fuera de su control, la filosofía Open no ha llegado todavía a ese punto de penetración y respecto a la navegabilidad y descubrimiento de recursos se sabe que los recursos MARC no son navegables fuera de la biblioteca y está por demostrar que los recursos Linked Data si lo sean.
7. Que las grandes compañías como Google entren en el mundo de los datos vinculados sólo se producirá si existe una perspectiva económica para ello. Por ello desde algunos foros se sugiere que se provean estímulos públicos para aumentar el ámbito de utilización de estas tecnologías. Las posibilidades de Linked Data pueden aparecer en el medio plazo, pero tal y como afirma Christensen, no hay mercado, sólo proyectos de investigación, además las bibliotecas pueden pensar que sus datos bibliográficos no son necesarios al menos con un grado fino de granularidad, pues las necesidades de los usuarios en cuanto a metadatos bibliográficos están más orientadas a los datos básicos (Moulaison & Million, 2014).

Existe una gran preocupación con respecto a la calidad de los datos, no sólo los actuales, sino también los futuros. En no pocas ocasiones quienes publican los datos, deben realizar tareas de limpieza de los datos y de “reconciliación” de los mismos (proceso por el cual se asigna a una cadena de texto la relación semántica exacta con su concepto). Para la realización de estas tareas se requieren grados elevados de competencias en técnicas y manejo fluido de vocabularios semánticos. La edición manual de estos dataset se convierte pues en un obstáculo tanto a la publicación, como a la vinculación pasando por el mantenimiento de los conjuntos de datos. A día de hoy, las bibliotecas no están preparadas para esto, se requiere el desarrollo de más automatismos y aplicaciones. (Van-Hooland et al., 2012).

Respecto a la recuperación de la información, no existe ningún producto semántico que pueda presentarse como una plataforma generalista de búsqueda efectiva. La búsqueda en catálogos está evolucionando hacia el usuario, a satisfacer sus necesidades de información y a mejorar su experiencia de búsqueda en catálogos. Estas tareas se han visto en parte satisfechas con la mera utilización de palabras clave y a los sumo algún operador booleano. No está claro que Linked Data facilite un proceso semejante al anterior, y aunque es cierto que SPARQL permite la confección de complejas consultas con resultados pertinentes, no parece fácil que el usuario admita estas tecnologías de gran complejidad, que, a no ser que estén intermediadas por interfaces que

faciliten su usabilidad, no parece fácil que vayan a implantarse como una solución estándar de búsqueda. (Moulaison & Million, 2014; Saorín et al., 2013) .

Lo que también parece claro es que los problemas que presenta Linked Data son solucionables y en ese sentido podemos apuntar algunas ideas:

1. Migración hacia sistemas de gestión de bibliotecas más ágiles que se apoyen en las bondades que pueden ofrecer los datos vinculados.
2. La biblioteca se puede presentar como principal adalid de estas tecnologías, aportando la garantía de su buen hacer por años a este nuevo marco.
3. Fomentar la cooperación transversal y aportar la experiencia en la gestión de datos internos y su preservación.
4. Evolucionar en las resistencias internas a la publicación de datos, mentalizando de la necesidad de exportar los datos bibliotecarios y con ello recuperar el protagonismo de los centros de información en la gestión de la misma.
5. Promover planes de formación, también al nivel de los gerentes de biblioteca.
6. Ser promotores de la descripción de procedencia, de la asignación de metadatos a los datasets, de la necesidad de establecer parámetros de calidad y fiabilidad de los datos y la preservación de los datos con esas mismas condiciones.

En definitiva, se requiere un cambio de perspectiva, los propios centros pueden reconvertirse en impulsores de proyectos como la colaboración con empresas tecnológicas. Los sectores científicos y de investigación están muy interesados en el desarrollo de tecnologías de gestión de datos y la colaboración con ellos puede aportar a las bibliotecas los recursos necesarios para liderar la evolución durante la actual de transición de los sistemas de información y como resultado de todo ello estar mejor posicionados respecto a sus competidores por la información. (Moulaison & Million, 2014).

4 REPRESENTACIÓN DE SISTEMAS DE ORGANIZACIÓN DEL CONOCIMIENTO

4.1 SISTEMAS PARA LA ORGANIZACIÓN DEL CONOCIMIENTO

Tradicionalmente las bibliotecas han controlado el lenguaje que utilizaban para aplicarlo en sus tareas habituales de indización y recuperación documental. Estos sistemas de organización del conocimiento (KOS) han evolucionado al compás de las nuevas tecnologías, ofreciendo nuevas perspectivas de utilidad y conviviendo con nuevos modos de organizar el conocimiento a través de vocabularios muy adaptados a los entornos digitales. Los tesauros han migrado hasta su versión conceptual, las taxonomías han cobrado un nuevo impulso para ordenar el conocimiento en el sector privado, fundamentalmente en entornos Web, las listas de encabezamientos de materia que lideran la publicación de sus estructuras en Linked Data. Frente a ellos están los KOS nacidos dentro del contexto digital, cuya principal característica es la de mimetizarse de modo natural con el entorno web. Es el caso de las folksonomías producidas por la cooperación no organizada en Internet, o las redes semánticas de conceptos para la representación de recursos digitales complejos, o los sistemas relacionales de representación del conocimiento mediante grafos, como los mapas conceptuales e incluso los Topic maps. (International Organization for Standardization (ISO), 2006).

Los vocabularios controlados de cuño bibliotecario tienen un papel protagonista que representar. Aportan al mundo de la información sus ventajas y lo mejor de la vieja escuela de la organización del conocimiento, añadiendo a sus tradicionales ventajas para las funciones de indización y recuperación de la información, las de integrarse en la Web y por ello aumentar exponencialmente su utilidad. Podemos hablar de un sistema que conforma una plataforma de servicios interconectados, donde los vocabularios, los sistemas de recuperación y almacenamiento de datos, ofrecen un modo sencillo de estructurar y visualizar el conocimiento (Méndez-Rodríguez & Greenberg, 2012; Moreira-Gonzalez, 2011).

El modo en el que estos KOS establecen su funcionamiento supone el establecimiento de relaciones entre conceptos y recursos, configurando una red semántica formada por nodos de recursos interconectados. En referencia a ello, algunos autores apuntan la necesidad de considerar a los documentos como una extensión del concepto, lo que permitiría por ejemplo que la visualización de conceptos asociados mostrará en realidad campos semánticos integrados por recursos (Baker et al., 2013).

Como se ha referido anteriormente, las listas de encabezamientos de materia, han sido pioneras en la publicación en Linked Data, como las *Subject Headings* de la Library of Congress, los vocabularios de materia RAMEAU de la Biblioteca Nacional Francesa o el Servicio de Datos

vinculados de autoridad de la Biblioteca Nacional de Alemania (DNB). Habitualmente se han codificado con vocabularios específicos para KOS que describen sus propiedades y estructura, y son identificadas necesariamente, mediante IRI que permite su trasposición a la nube de datos. Esta migración desde un literal a un IRI enlazable es la esencia de la conversión de los encabezamientos de materia y otros sistemas de organización del conocimiento. Por ejemplo: <http://id.loc.gov/authorities/subjects/sh92004914>, es un IRI que representa la materia *Semantic network* en la lista de encabezamientos de materia en Linked Data de la Library of Congress y que, a diferencia de un literal, permite identificar el valor (materia asignada) y reutilizarla libremente en la nube de datos (Isaac, Waites, Young, & Zeng, 2011).

A continuación vamos a hacer una breve referencia a los KOS implicados en la realización de este proyecto: tesauros y encabezamientos de materia.

4.1.1 TESAUROS Y ENCABEZAMIENTOS DE MATERIA. KOS PARA LA WEB DE DATOS

4.1.1.1 Encabezamientos de materia

Una lista de encabezamientos de materia es un tipo de vocabulario controlado que representa de modo sintetizado temas contenidos en un documento de cualquier tipo. Al igual que los tesauros se compone de conceptos en forma de términos o frases y del mismo modo que los esquemas de clasificación utiliza reglas sintácticas para combinar términos en secuencias precoordiadas que representan conceptos más complejos. Los encabezamientos de materia recogen tópicos sobre colecciones de información lo que permite organizarlas de modo sistemático en función de su contenido y facilitar la navegación sobre el dominio de la materia. Su proceso de construcción lógico es complejo, pues en el momento de la indización, el catalogador escoge componentes necesarios (uno o varios) para representar las facetas de la materia. (Gil-Urdiciain, 2004; International Organization for Standardization (ISO), 2012; Moreira-Gonzalez, 2011).

Los principales componentes de una lista de encabezamientos de materia son los encabezamientos, las subdivisiones (subencabezamientos) y las relaciones entre ellos.

Las formas en que se presentan los encabezamientos son tres:

1. Una palabra que representa un concepto.
2. Varias palabras que representan un concepto.
3. Varias palabras que representan una combinación de varios conceptos.

Lo más característico de las listas de materias son las formas complejas en las que al encabezamiento inicial se le añaden uno o más encabezamientos secundarios o subdivisiones para representar un concepto de modo completo y preciso. Estos subencabezamientos

representan un punto de vista, o forma bajo el cual el encabezamiento presenta la información del recurso en las típicas subdivisiones temáticas, geográficas, cronológicas y de forma.

Respecto al control del vocabulario, los homógrafos se desambiguan con un calificador o adición. Los encabezamientos pueden ser aceptados (por el centro catalogador) o no aceptados, los primeros son asimilables a los términos preferidos, los segundos no son válidos para la indización. Sus relaciones habituales son las de equivalencia y asociación, aunque cabe también la utilización de relaciones jerárquicas.

Su representación en Linked Data no se aleja apenas a la de los tesauros, la mayor complejidad es la determinada por los encabezamientos complejos precoordinaados, que se describen primariamente como conceptos únicos y en ocasiones, mediante vocabularios u ontologías especiales que lo descomponen.(Gil-Urdiciain, 2004; International Organization for Standardization (ISO), 2012; Moreiro-Gonzalez, 2011)

4.1.1.2 Tesauros

El modelo ISO 25964 va a ser el utilizado en este proyecto como base para la publicación de las materias de la BUPM en Linked Data. La norma ha evolucionado para actualizar la gestión de vocabularios en un contexto distribuido. Su parte primera define los parámetros en base al trabajo intelectual con el léxico para mejorar la indización y la recuperabilidad, sin perder de vista las recomendaciones sobre formatos de intercambio y protocolos ni la definición de un modelo de datos orientado a dar soporte a vocabularios web. La parte segunda estructura y define los complejos sistemas de alineamiento y mapeo entre vocabularios y sus conceptos, ofreciendo un completo marco de interoperabilidad definido para una amplia variedad de tipos de vocabularios. Todo ello conforma un marco complejo cuyos caracteres básicos son la participación humana en la concepción intelectual del vocabulario, la interoperabilidad entre vocabularios, la gestión automatizada de tesauros y la compatibilidad con aplicaciones, fundamentalmente bases de datos. La propia norma se habilita para la definición de sistemas típicamente post coordinados, pero también lo hace para otros muchos tipos de vocabularios, entre ellos los de naturaleza pre-coordinada, como las listas de encabezamientos de materia (International Organization for Standardization (ISO), 2011).

La estructura básica de la norma desde un punto de vista técnico puede resumirse del siguiente modo:

1. Modelo de datos orientado a las aplicaciones informáticas y a la recuperación de la información.
2. Estructura conceptual del tesoro representada léxicamente.
3. Definición de diversos niveles de agrupación semántica de conceptos.
4. Relaciones de equivalencia con varios niveles de complejidad.

A los efectos de la Web de datos, un tesoro es un vocabulario integrado por conceptos relacionados y referidos a un dominio definido al que organiza (International Organization for Standardization (ISO), 2011). Los conceptos están representados por términos o etiquetas. Su expresión sintáctica puede establecerse a través de una estructura simple (un sólo concepto) o múltiple (varios conceptos), con la posibilidad de hacer combinaciones libres durante la indización o la recuperación. Los términos son representaciones (etiquetas) de los conceptos, por ello, un concepto puede ser representado por varios términos (incluso una combinación de ellos, o un calificador), pero un término sólo designa un concepto. Uno de los términos representativos, el que mejor defina al concepto para la comunidad que vaya a utilizar el tesoro, debe ser especificado como preferente, los demás serán considerados no preferentes y referenciados al primero. Los conceptos complejos son aquellos formados por varios conceptos con diferente nivel de importancia en la expresión. Pueden estar formados por multitérminos o términos compuestos.

El sistema de relaciones que define la norma ISO 25964 es el paradigmático, recogiendo las típicas relaciones de equivalencia, jerárquicas y asociativas. (Keyser & Leuven, 2012; Moreira-Gonzalez, 2011). A estas se suman algunas específicas, como la equivalencia compuesta, que explicaremos en el contexto de los mapeos, o ciertas relaciones especiales establecidas entre un concepto y su acrónimo o un concepto y su *TopConcept*.

El sistema de agrupación de conceptos regulado por la norma permite el diseño de colecciones o *arrays*, y la creación de microtesoros (etiqueta *ConceptGroup*). Los conceptos incluidos en el microtesoro pueden tener o no relaciones jerárquicas entre sí y se agrupan en torno a un núcleo semántico que les engloba. El modelo de datos de la norma ISO dispone de clases que identifican a los grupos y sus etiquetas y permite que se aniden unos grupos de conceptos dentro de otros jerárquicamente. Los conceptos del microtesoro pueden provenir de diferentes jerarquías (polijerarquías) y facetas del tesoro (International Organization for Standardization (ISO), 2011).

Finalmente la norma contempla un sistema completo de anotación que permite documentar diferentes aspectos de sus conceptos o términos, su alcance, su justificación como término preferente, la fuente de donde se seleccionó, etc.

4.1.2 VOCABULARIOS DE VALORES

En este trabajo se va a efectuar un modelado semántico y publicación de un vocabulario de valores, concretamente la lista de encabezamiento de materias de la BUPM. Los vocabularios de valores no son ajenos al fenómeno Linked Data, son muchas las publicaciones que se han efectuado y muchas de ellas ejemplo de calidad en todos sus procesos. Como se puede observar

en el listado siguiente, los principales vocabularios de referencias ya tienen su plasmación en la Web de datos, sin ánimos de ser exhaustivos, referimos los siguientes:

- a. El sistema de clasificación Dewey (OCLC, 2014b) y CDU (UDC Consortium, 2014).
- b. La colección de datasets y vocabularios de ámbito bibliotecario de la Library of Congress (Library of Congress, 2014b).
- c. El repertorio de autoridades de materia de la Biblioteca Nacional de Francia, “RAMEAU” (Bibliothèque nationale de France, 2014b).
- d. El servicio Linked Data de la Biblioteca Nacional de Alemania (Deutsche Nationalbibliothek, 2014b).
- e. La lista de encabezamientos de materia de las bibliotecas públicas de España (Ministerio de Educación, Cultura y Deporte, 2014).
- f. El Directorio de ficheros de autoridad VIAF donde se recogen datos de autoridad de lugares, obras, entidades, personas, títulos y más (OCLC, 2014c).
- g. La base de datos de nombres geográficos GEONAMES (Geonames, 2014).
- h. AGROVOC es un vocabulario controlado de interés para la FAO (*Food and Agriculture Organization*) (Food and Agriculture Organization of the United Nations, 2014).
- i. Los vocabularios de licencias Creative Commons (Creative Commons, 2014a):
- j. Base del conocimiento Dbpedia (University of Mannheim, OPENLIK, & Universität Leipzig, 2014).

4.2 SIMPLE KNOWLEDGE ORGANIZATIONN SYSTEM: SKOS

La necesidad de migrar KOS al espacio de datos vinculados motivó el desarrollo de un lenguaje ontológico de amplio espectro cuyo rango de descripción abarcara el máximo número de supuestos de representación. SKOS ofrece un espacio de modelado generalista de conceptos y sus relaciones, siendo precisamente esta característica la que describe a la vez la fortaleza y la debilidad de su sistema, pues su capacidad de descripción de los rasgos comunes de una amplia tipología de KOS, se ve debilitada por carencias en el nivel de granularidad que se puede llegar a conseguir. Sus desarrolladores prefirieron que SKOS perdiera precisión de modelado para evitar que los motores de inferencia obtuvieran conclusiones erróneas. Su objetivo era más obtener un estándar de amplia utilización, permitiendo un mayor ajuste mediante la utilización de subclases o subpropiedades, o el empleo de vocabularios de extensión como puede ser Dublín Core. A continuación se va a realizar un breve acercamiento al estándar; por un lado se abordará la teoría general del modelo de datos de SKOS, por otro se muestra el desarrollo del modelo con ejemplos y se describe la extensión SKOS-XL cómo ayuda a la gestión de los vocabularios de valores y finalmente se definen las debilidades y posibilidades de mejora del estándar.

4.2.1 ESTRUCTURA GENERAL DEL MODELO SKOS

El modelo semántico de SKOS supera en compatibilidad al estándar OWL para la representación de dominios concretos como ontologías formales. OWL específica, con un gran nivel de definición, la estructura del conocimiento en dominios concretos, mientras que SKOS ofrece estructuras generales descritas con pocas reglas de modelado.

El núcleo del modelo SKOS separa espacios de representación, lo que doctrinalmente se denomina disjunción de dominios. Por un lado establece la idea de concepto (asumida también en la nueva norma de tesauros ISO 25964) que se separa estructuralmente de otros ámbitos de agrupación como `skos:Collection`, `skos:ConceptScheme`, `skosxl:Label`. Esa estructura general asigna una posición a cada concepto en SKOS, como miembro de un esquema general (`skos:inScheme`) y a su vez como integrante de la estructura jerárquica conceptual del “*Scheme*”, ya en una posición de cabecera semántica (`skos:hasTopConcept`), o como integrante de la cadena jerárquica general (`skos:Concept`). En cambio, `skos:Collection` no se configura como un sistema de agrupación conceptual al mismo nivel sintáctico que `skos:ConceptScheme` y no puede establecer relaciones semánticas como ente propio, con otras estructuras del modelado SKOS, aunque puede contener conceptos agrupados por campos semánticos, pero sin unidad estructural relacionable. Esta característica es muy importante para la representación de dominios específicos, por ejemplo los microtesauros o las agrupaciones de subdivisiones de materia en `skos:Collection` (EUROVOC, utiliza microtesauros como

“conceptScheme” la Library of Congress, en ocasiones, utiliza colecciones para sus subencabezamientos de materia (Baker et al., 2013).

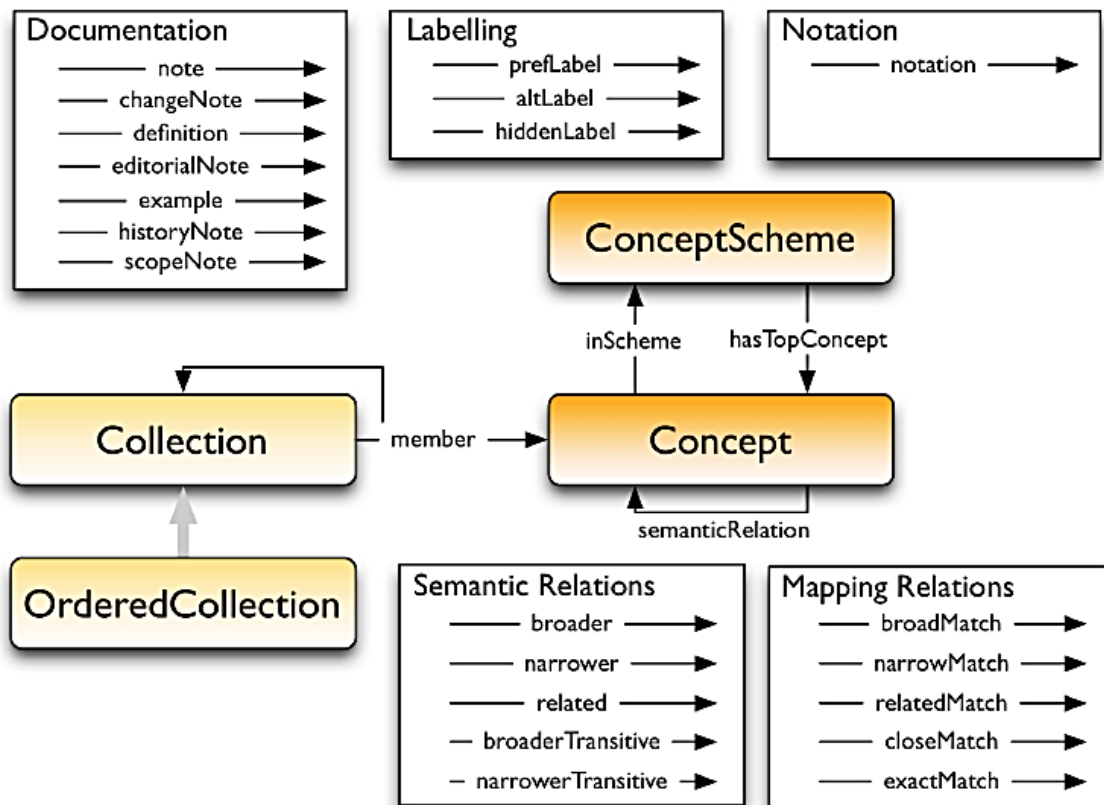


Figura 18 Elementos básicos del modelo de datos SKOS (Baker et al., 2013)

4.2.2 ESQUEMA DEL MODELO DE DATOS DE SKOS

Simple Knowledge Organization System (SKOS) es un lenguaje derivado de RDF para la representación de esquemas de conceptos como taxonomías, esquemas de clasificación, listas de encabezamientos de materias, tesauros etc., en el contexto de Linked Data. El objetivo de SKOS es la representación de vocabularios conceptuales en la Web de datos, permitiendo la reutilización y la interoperabilidad entre dichos esquemas. Podemos enumerar las características fundamentales de su estructura de modelado.

4.2.2.1 Conceptos

El elemento primario de SKOS es el concepto o idea que queremos estructurar y cuya propiedad es `skos:Concept`. Desde el punto de vista de modelado de SKOS, un concepto puede aparecer en diferentes esquemas dentro del mismo KOS. Un sencillo ejemplo puede ser:

```
<http://www.upm.es/biblioteca/kos/sh/Programación> a skos:Concept .
```

4.2.2.2 Etiquetas léxicas

Para la expresión en lenguaje natural de los conceptos, SKOS emplea etiquetas que se asimilan a términos referidos en las normas de tesauros. Se percibe aquí un cierto alineamiento de la ISO 25964, que ha adoptado también el modelo conceptual de tesoro y que concibe a los términos como representaciones o etiquetas de los conceptos. La etiqueta preferente para un recurso se expresa mediante `skos:prefLabel`, no pudiendo haber dos conceptos con la misma etiqueta preferente en un mismo idioma. La propiedad `skos:altLabel` permite asignar etiquetas alternativas a los conceptos, es decir términos no preferidos. Con la propiedad `skos:hiddenLabel` se establece un sistema que permite ocultar etiquetas en la interfaz humana, para no mostrar formas incorrectas léxicamente de los términos, pero que en la estructura técnica del vocabulario tienen el valor de etiqueta o término alternativo y por tanto son hábiles para la recuperación:

```
<http://www.upm.es/biblioteca/kos/sh/programación informática> a skos:Concept;  
  skos:prefLabel "programación informática";  
  skos:altLabel "serialización" .
```

4.2.2.3 Relaciones entre conceptos

Las relaciones son un elemento central en el marco de los datos vinculados. Los conceptos están relacionados de tres modos fundamentales en SKOS. Para la representación jerárquica se utilizan las propiedades `skos:broader`, el objeto de la tripleta es un concepto más general o de mayor alcance que el concepto que es sujeto; `skos:narrower`, el objeto es más específico que el sujeto de la tripleta y para las relaciones asociativas se utiliza la propiedad `skos:related`, por ejemplo:

```
<http://www.upm.es/biblioteca/kos/sh/lenguajes de programación> a  
skos:Concept;  
  skos:prefLabel "lenguajes de programación";  
  skos:narrower "lenguaje C";  
  skos:broader "programación";  
  skos:related "lingüística informática" .
```

Las relaciones jerárquicas referenciadas son intransitivas, es decir no se puede deducir en una jerarquía de dos niveles que un término genérico A, respecto a otro específico B, no transmite su

relación de generalidad a los conceptos específicos de B. Para mejorar la inferencia fundamentalmente en el ámbito de las aplicaciones automáticas, SKOS dispone de las propiedades `skos:broaderTransitive` and `skos:narrowerTransitive` que permiten extender la semántica más allá de dos niveles de jerarquía. Podemos establecer la siguiente inferencia según el ejemplo siguiente: si establecemos una propiedad transitiva en la relación entre Redes y Redes locales, y entre Redes locales e Intranets, podemos asegurar que Intranets es un concepto específico de Redes. Esta deducción de tan obvia expresión, no es tal en las relaciones genérico-específico del estándar SKOS, donde ese conocimiento no se puede deducir.

```
<http://www.upm.es/biblioteca/kos/sh/Redes> a skos:Concept;  
  skos:prefLabel "Redes".  
  
<http://www.upm.es/biblioteca/kos/sh/Redes locales> a skos:Concept;  
  skos:prefLabel "Redes locales";  
  skos:broader "Redes" .  
  
<http://www.upm.es/biblioteca/kos/sh/Intranets> a skos:Concept;  
  skos:prefLabel "Intranets";  
  skos:broader: "Redes locales" .
```

Si se utilizan las propiedades transitivas, las aplicaciones pueden deducir que Redes es un conceptos genérico respecto a Redes locales que transmite esa relación a niveles más bajos de la jerarquía. Lo mismo podemos decir del conceptos Redes locales.

```
<http://www.upm.es/biblioteca/kos/sh/Redes locales> a skos:Concept;  
  skos:prefLabel "Redes locales";  
  skos:broaderTransitive "Redes";  
  skos:related "Redes Wi-Fi" .  
  
<http://www.upm.es/biblioteca/kos/sh/Intranets> a skos:Concept;  
  skos:prefLabel "Intranets";  
  skos:broaderTransitive "Redes locales" .
```

Por lo tanto, y en este caso, el modelado mediante `skos:broadertransitive` permite deducir que Redes es genérico también en el segundo escalón de la jerarquía respecto de Intranets.

```
<http://www.upm.es/biblioteca/kos/sh/Intranets> a skos:Concept;  
  skos:prefLabel "Intranets";  
  skos:broaderTransitive "Redes" .
```

4.2.2.4 Anotaciones

En ocasiones la semántica del concepto debe ser especificada para completar su sentido. SKOS permite hacer anotaciones sobre los conceptos en lenguaje natural y comprensible por los humanos. Son varias las posibilidades para documentar los conceptos que ofrece SKOS, las más importantes son:

1. `skos:note`, para informaciones sobre el concepto de carácter general.
2. `skos:scopeNote`, ofrece una breve nota de información sobre el concepto referenciada especialmente a su contexto de uso.
3. `skos:definition`, facilita una explicación extensa y completa del concepto.
4. `skos:historyNote`, describe cambios en el significado del concepto a través del tiempo.
5. `skos:changeNote`, ofrece información sobre cambios específicos en el concepto.
6. `skos:editorialNote`, incluye información relevante para los editores de vocabularios sobre cambios, actualizaciones, etc.

```
<http://www.upm.es/biblioteca/kos/sh/script> a skos:Concept;  
    skos:prefLabel "script";  
    skos:note "archivo de procesamiento por lotes";  
    skos:definition "es un programa usualmente simple, almacenado en un  
    archivo de texto plano. Sus instrucciones son procesadas por el  
    intérprete de lenguajes de programación, ejecutando cualquier tarea  
    definida en su guion. Su intermediario de actuación es el Shell" .
```

4.2.2.5 Agrupaciones semánticas de conceptos

Habitualmente, los sistemas de organización del conocimiento se despliegan sobre un dominio determinado que a su vez puede estructurarse en campos semánticos más pequeños designados mediante `skos:ConceptScheme`. Los conceptos que integran los esquemas declaran su pertenencia a los mismos mediante `skos:inScheme`, pudiendo cada concepto pertenecer a uno o varios esquemas.

```
<http://www.upm.es/biblioteca/kos/sh/BUPMSubjectHeadings> a  
    skos:ConceptScheme .  
  
<http://www.upm.es/biblioteca/kos/sh/lenguaje de programación> a  
    skos:Concept;  
    skos:inScheme "BUPMSubjectHeadings";  
    skos:narrower "lenguaje C";  
    skos:broader "programación";  
    skos:related "lingüística informática" .
```

El esquema contiene conceptos principales que ocupan la cabecera de las jerarquías, son los denominados *Top Concepts*. Cada esquema debe tener al menos un concepto de cabecera declarado mediante `skos:hasTopConcept`. Los conceptos de cabecera a su vez pueden

declarar su pertenencia a un *ConceptScheme* declarando la propiedad `skos:topConceptOf` (se utiliza en el ejemplo el espacio de nombres DCMI terms):

```
<http://www.upm.es/biblioteca/kos/sh/BUPMSubjectHeadings> a
skos:ConceptScheme;
    dct:title "Encabezamientos material Biblioteca UPM";
    skos:hasTopConcept "Telecomunicaciones";
    skos:hasTopConcept "Informática" .

<http://www.upm.es/biblioteca/kos/sh/Telecomunicaciones> a skos:Concept;
    skos:inScheme "BUPMSubjectHeadings";
    skos:topConceptOf
<http://www.upm.es/biblioteca/kos/sh/BUPMSubjectHeadings> .
```

SKOS permite la agrupación de conceptos a través de la clase `skos:collection`. La norma ISO 25964 hace referencia a las colecciones de conceptos como “*arrays*” que están identificados mediante una etiqueta de nodo. La pertenencia de un concepto a una colección se declara mediante la instancia `skos:member`.

```
<http://www.upm.es/biblioteca/kos/sh/LenguajesC> a skos:collection;
skos:prefLabel "Lenguajes de programación en C";
skos:member <http://www.upm.es/biblioteca/kos/sh/Lenguajes C+>;
skos:member <http://www.upm.es/biblioteca/kos/sh/Lenguajes C++>;
skos:member <http://www.upm.es/biblioteca/kos/sh/Lenguajes C#> .
```

En ocasiones es necesario ordenar mediante algún criterio las listas de conceptos de las colecciones, para ello SKOS utiliza la clase `skos:OrderedCollection`. En este caso la pertenencia a una colección ordenada se declara mediante la instancia `skos:memberlist`.

```
<http://www.upm.es/biblioteca/kos/sh/Lenguajes C> a skos:OrderedCollection;
skos:prefLabel "Windows (Sistemas Operativos)";
skos:memberList <http://www.upm.es/biblioteca/kos/sh/Windows 3.1> ;
skos:memberList <http://www.upm.es/biblioteca/kos/sh/Windows 95> ;
skos:memberList <http://www.upm.es/biblioteca/kos/sh/Windows XP> ;
skos:memberList <http://www.upm.es/biblioteca/kos/sh/Windows 7> .
```

4.2.3 SKOS EXTENSION FOR LABELS. SKOS-XL.

La extensión de SKOS para etiquetas posibilita la identificación de relaciones entre entidades léxicas (etiquetas). SKOS-XL concibe las etiquetas como recursos que pueden ser objeto de una declaración RDF y por ello relacionables.

En primer lugar `skosxl:Label` es una clase especial de entidad léxica, cuya instancia puede ser un recurso identificado por un IRI. Una instancia de `skosxl:label` puede ser un literal (sólo un literal por cada `skosxl:label`), esa expresión literal se materializa a través de la propiedad `skosxl:literalForm`. La propiedad `skosxl:labelRelation` enlaza sendas entidades léxicas de `skosxl:Label` siendo una propiedad extensible que puede utilizarse para definir enlaces más específicos. Las diferentes entidades de `skosxl:Label` pueden declararse como integrantes de un esquema de conceptos dado, esta pertenencia se declara mediante `skos:inScheme`.

Como hemos dicho, los IRIs deben ser opacos para evitar problemas de identificación; podemos establecer un sistema para vincular el código de la materia con la versión legible por humanos de la misma:

```
<http://www.upm.es/biblioteca/kos/sh/XX2358> a skos:Concept ;
skosxl:prefLabel <http://ex.upm.es/biblioteca/sh/etiquetas/InternetOfThings> ;
<http://ex.upm.es/biblioteca/sh/etiquetas/InternetOfThings> a skosxl:Label ;
skosxl:literalForm "Internet of Things" ;
skos:inScheme <http://www.upm.es/biblioteca/kos/sh/BUPMSubjectHeadings> .
```

Las equivalencias entre entidades léxicas se expresan en SKOS-XL de modo muy similar a SKOS. La propiedad `skosxl:prefLabel` asigna etiquetas preferentes a los recursos, `skosxl:altLabel`, asigna etiquetas no preferentes y `skosxl:hiddenLabel`, oculta la etiqueta en la interfaz humana manteniéndola accesible para lectura automática. Todas estas propiedades son instancias de `skosxl:Label`. (Miles & Bechhofer, 2009). En el ejemplo inferior se muestra las diferentes posibilidades de jerarquías léxicas en SKOS-XL:

```
<http://www.upm.es/biblioteca/kos/sh/XX5685> a skos:Concept ;
skosxl:prefLabel http://ex.upm.es/biblioteca/sh/etiquetas/LinkedData ;
<http://ex.upm.es/biblioteca/sh/etiquetas/LinkedData> a skosxl:Label ;
skosxl:literalForm "Linked Data" ;
skosxl:altLabel http://ex.upm.es/biblioteca/sh/etiquetas/LD ;
<http://ex.upm.es/biblioteca/sh/etiquetas/LD> a skosxl:Label ;
skosxl:literalForm "LD" .
```

Una de los principales problemas de expresividad semántica de los encabezamientos de materia en Linked Data es la pérdida del proceso de precoordinación de encabezamientos. Una posible

solución usando el estándar SKOS y SKOS-XL como base sería construir un sistema que permitiera enlazar conceptos vinculados.

Se genera un modelo de ejemplo con dos conceptos: “Linked Data” y “Sistemas de información”, concepto que se pretende vincular semánticamente en un encabezamiento pre coordinado, donde actúa como subencabezamiento temático.

Se crean sendas declaraciones que identifican los conceptos y sus etiquetas preferentes, dichas etiquetas se declaran como recursos mediante el modelado SKOS-XL.

```
<http://www.upm.es/biblioteca/kos/sh/XX5685> a skos:Concept ;
skosxl:prefLabel http://ex.upm.es/biblioteca/kos/sh/etiquetas/LinkedData ;
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/LinkedData> a skosxl:Label ;
skosxl:literalForm "Linked Data" ;
skosxl:altLabel http://ex.upm.es/biblioteca/kos/sh/etiquetas/LD ;
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/LD> a skosxl:Label ;
skosxl:literalForm "LD" .
```

```
<http://www.upm.es/biblioteca/kos/sh/XX5869> a skos:Concept ;
skosxl:prefLabel
http://ex.upm.es/biblioteca/kos/sh/etiquetas/SistemasdeInformación ;
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/SistemasdeInformación> a
skosxl:Label ;
skosxl:literalForm "Sistemas de Información" ;
skosxl:altLabel http://ex.upm.es/biblioteca/kos/sh/etiquetas/LD ;
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/SistemasInformativos> a
skosxl:Label ;
skosxl:literalForm "Sistemas informativos" .
```

Tal y como aparece en el estándar (Apéndice B, SKOS Reference 2009) podemos vincular ambas etiquetas pero no se identificaría exactamente el tipo de relación, únicamente que una está en relación con la otra, mientras que lo que se pretende declarar es que “Sistemas de Información” es subencabezamiento temático de “Linked Data”.

```
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/LinkedData>
skosxl:labelRelation
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/SistemasdeInformación> .
```

Para conseguir la expresividad semántica completa, podemos declarar que los subencabezamientos son subpropiedades de `skosxl:labelRelation`. Para ello se crea un “*namespace*” que contenga como propiedades los tipos de encabezamiento, y se declara su subordinación (Miles & Bechofer, 2009; Miles & Bechhofer, 2009).


```
<http://ex.upm.es/biblioteca/kos/sh/subenc/IsTopicalSubdivisionOf>
#prefix:subenc# rdfs:subPropertyOf skosxl:labelRelation .

<http://ex.upm.es/biblioteca/kos/sh/etiquetas/LinkedData>
subenc:IsTopicalSubdivision
<http://ex.upm.es/biblioteca/kos/sh/etiquetas/SistemasdeInformación> .
```

4.2.4 EVOLUCIÓN DEL MODELO SKOS

El problema de la precoordination de conceptos en los KOS es un punto de resolución pendiente en SKOS. Un concepto (encabezamiento de materia) puede ser especificado mediante otro concepto (subencabezamiento de materia) mediante la suma de ambos de modo coyuntural, conformando un nuevo concepto. La justificación de esta carencia se basa en que en el ámbito de los tesauros, dos conceptos simples pueden unirse en uno nuevo compuesto en el propio momento de la recuperación, mientras que en otros KOS lo hace con carácter previo.

Una posible solución no integrada aún en la estructura de modelado SKOS, consiste en la relación de conceptos simples mediante operadores booleanos; otra es la que establece la propia recomendación del W3C en su texto, la integración de propiedades en las declaraciones. Entre tanto la Library of Congress ha desarrollado una versión de MADS/RDF que permite entre otras cosas introducir la precoordination en los encabezamientos de materia.

En el ejemplo siguiente el W3C da una propuesta formal que no es normativa (a modo de ejemplo se utiliza un *namespace* no real que incluye las propiedades necesarias):

```
shupmp:coordinationOf a rdf:Property;
  rdfs:domain skos:Concept;
  ex:Arte S. XII-XV a skos:Concept;
  shupmp:coordinationOf (shupmp:Arte shupmp:S. XII-XV);
  skos:prefLabel "Arte -- S. XII-XV" .
```

En otro orden de cosas, en el conjunto de desarrolladores de sistemas de recuperación e indización se está percibiendo la necesidad de establecer una vinculación semántica estructurada entre los conceptos y los recursos que son indizados o recuperados por ellos. Se trata de conseguir un corpus documental definido mediante la especificación del concepto que los une en un campo semántico.

La cuestión de la reciprocidad en las relaciones entre elementos modelados con SKOS, también requiere una adaptación del estándar. SKOS no contempla las relaciones inversas a la hora de aplicar sus elementos `skos:inScheme` o `skos:member`, los elementos vinculados en una dirección no “devuelven” la relación mediante ningún elemento SKOS. La aplicación de marcado semántico con RDFa permite superar esas carencias semánticas a falta de las esperadas

extensiones de SKOS. Tampoco define SKOS los puntos de entrada a la consulta de conjuntos de conceptos, cuestión relevante si se construyen por ejemplo vocabularios de colecciones de conceptos o microtesaruros (Pastor-Sánchez, 2013b).

También en este ámbito es imprescindible añadir a los datasets, descripciones de metadatos que ayuden a su identificación. Tan importante es el contenido intelectual de los sistemas como la descripción adecuada de los mismos para hacerlos visibles. Por ello se pide la definición de nuevas posibilidades de descripción incluidas en el mismo modelo de SKOS, sin acudir a vocabularios embebidos. Concretamente se pide la inclusión obligatoria de información descriptiva sobre los “*ConceptSchemes*”, descripción de la evolución cronológica de los conceptos, de información de procedencia sobre los mapeos que se establecen entre vocabularios y sobre licencias de datos (Baker et al., 2013).

4.2.5 MAPEOS CON SKOS

A efectos del prototipo de lista de encabezamientos de materia que se diseña en este proyecto, conviene hacer alusión a algunas características del modelado SKOS para el mapeo de vocabularios en el contexto de la Web de datos.

El mapeado pone en relación conceptos de dos o más vocabularios a través de su proximidad semántica, de hecho el mapeado de conceptos es la base de construcción de redes de conceptos en la Web semántica. La forma correcta de establecer un mapeado entre diferentes esquemas de conceptos es el diseño de los mismos como IRIs. Para especificar que dos conceptos o más, de diferentes esquemas tienen un significado similar, SKOS provee las propiedades `skos:exactMatch` y `skos:closeMatch`, la primera es subpropiedad de la segunda y se diferencian por la mayor similitud de los conceptos relacionados cuando se utiliza la primera frente a la segunda. En este ejemplo se vincula una materia primaria del prototipo frente a sendas materias de la Library of Congress y la Biblioteca Nacional Francesa.

```
<http://www.upm.es/biblioteca/kos/sh/Programación>  
skos:exactMatch  
<http://id.loc.gov/authorities/subjects/sh00007512>,  
<http://data.bnf.fr/ark:/12148/cb12042270g> .
```

También es posible el establecimiento de relaciones entre varios vocabularios en la que además de la vinculación por similitud semántica se especifican relaciones jerárquicas y lineales entre los conceptos a través de las propiedades derivadas de las anteriores: `skos:broadMatch`,

`skos:narrowMatch` y `skos:relatedMatch`. Una especificación importante es que SKOS no utiliza la propiedad `owl:sameAs` para las vinculaciones, esto sucede porque con ella los vocabularios quedan unidos como si fueran uno, frente a la utilización de `skos:exactMatch`, `skos:closeMatch` o `skos:relatedMatch` que suponen la vinculación únicamente del concepto.

SKOS permite la inclusión de un concepto de un vocabulario en otro para completarlo. Esta extensión de los vocabularios mediante los conceptos de otros es muy útil para la actualización y mejoras de KOS, utilizándose para ello propiedad `skos:inScheme`, que en este caso indica que el concepto sujeto se incluirá, a todos los efectos, en el esquema origen que actúa de objeto. Estamos ante un supuesto de interoperabilidad global, donde los vocabularios se generan escogiendo conceptos de los KOS disponibles en la Web de datos. Si en nuestro prototipo no aparece la materia “programación neurolingüística” podemos importarla desde la LC:

```
<http://id.loc.gov/authorities/subjects/sh85091129> skos:inScheme  
<http://www.upm.es/biblioteca/kos/sh/BUPMSubjectHeadings> .
```

SKOS no contempla en su modelo la vinculación concepto-recurso, ya hemos hablado que este tema es uno de los pendientes para posibles actualizaciones del estándar. En el marco de Linked Data, el enriquecimiento de la información es una de las más importantes ventajas. Parece lógico intentar vincular conceptos con sus documentos indizados, lo que supone que dichos documentos aparezcan en la estructura de elementos (Miles & Bechofer, 2009).

```
<http://www.upm.es/biblioteca/kos/sh/XX2358s> a skos:Concept;  
    skos:prefLabel "Internet of Things" .  
  
<http://www.bbc.com/future/story/20140413-why-ghosts-haunt-the-internet>  
a foaf:homepage;  
<http://purl.org/dc/terms/subject> "Internet of Things" .
```

4.3 AJUSTE DE LA NORMA ISO 25964 CON SKOS

La norma ISO 25964 permite un mayor ajuste de los vocabularios a la forma de representación de los KOS en la Web de datos. Como se ha dicho la norma ofrece un modelo de datos específico y con grandes similitudes respecto al modelo SKOS: modelo conceptual, sistemas de anotaciones, agrupaciones de conceptos, etc. A pesar de que es evidente que los creadores de la norma tuvieron en cuenta el estándar SKOS, existen algunas diferencias importantes, por lo que en la actualidad se está trabajando en el alineamiento de ambos modelos.

A día de hoy se está gestionando una tabla de correspondencias SKOS y SKOS-XL con las nueva norma ISO 25964, la extensión denominada ISO-THES (ISO TC46/SC9/WG8 working group for the ISO 25964 & Isaac, 2012). Se dispone ya del “*namespace*”, con fecha de septiembre de 2013, (Isaac & De Smedt, 2013) que permite una mejor estructuración y definición que el anterior esquema XML. Las equivalencias entre ambos espacios son lo suficientemente amplias como para no tener que definir una nueva ontología. Esto implica que la edición de un tesauro bajo ISO 25964 puede expresarse ya en Linked Data a través de la conversión a SKOS.

Algunas de las mejoras que aporta el nuevo esquema son la posibilidad de representar jerarquías de grupos de conceptos como los microtesauros, utilizados por ejemplo en el Tesauro de la UNESCO. También se abordan más eficientemente las equivalencias entre conceptos de un mismo vocabulario, especialmente las compuestas (International Organization for Standardization (ISO), 2011; Pastor-Sánchez, 2013a).

De gran importancia es el mapeo que se está realizando por el ISO TC46/SC9/WG8 working group for the ISO 25964, entre la norma ISO 25964, SKOS y SKOSXL. Se trata de ajustar las estructuras de ambos estándares y crear nuevos modelos de descripción para las lagunas en el alineamiento mediante propuestas de extensión de los vocabularios (ISO-THES).

En el modelo de datos que se definirá para la migración de la lista de encabezamientos de materia BUPM, se tendrá en cuenta los posibles paralelismos y lagunas que se verifiquen, proponiendo un modelo que recoja en lo posible los requerimientos de ambos estándares.

En su última versión de noviembre de 2013 el alineamiento básico respondía a la siguiente estructura (no se reflejan las correspondencias que están en fase de propuesta):

Tabla 5 Ajuste de los vocabularios ISO 25964 y SKOS (Isaac & De Smedt, 2013; ISO TC46/SC9/WG8 working group for the ISO 25964 & Isaac, 2012)

| Correspondencia entre ISO 25964 – SKOS - SKOSXL | |
|--|--|
| Clases generales | |
| Thesaurus | skos:ConceptScheme |
| ThesaurusConcept | skos:Concept |
| ThesaurusTerm | skos-xl:Label |
| isPartOf | skos:inScheme, skos:topConceptOf, skos:hasTopConcept |
| hasPreferredLabel | skos:prefLabel, skos-xl:prefLabel |
| hasNonPreferredLabel | skos:altLabel, skos-xl:altLabel, skos:hiddenLabel, skos-xl:hiddenLabel |
| Relaciones entre términos no representadas en SKOS, SKOSXL | |
| Agrupamientos | |
| hasMemberArray (puede no mantenerse el orden de conceptos) | skos:member |
| hasMemberConcept (puede no mantenerse el orden de conceptos) | skos:memberList |
| hasAsMember (puede no mantenerse el orden de los conceptos) | skos:member |
| Documentación | |
| Note | skos:note |
| ScopeNote hasScopeNote | skos:scopeNote |
| HistoryNote hasHistoryNote | skos:historyNote |
| Definition hasDefinition | skos:definition |
| CustomNote hasCustomNote | skos:changeNote, skos:example |
| Relaciones entre conceptos | |
| HierarchicalRelationship | skos:broader skos:narrower |
| AssociativeRelationship | skos:related |

4.4 MADS

MADS/RDF (Metadata Authority Description Schema in RDF) es una ontología, que se utiliza para la estructuración de la información en el contexto GLAM. A los efectos de este proyecto, MADS es un esquema que trabaja con vocabularios controlados y que tiene especiales características para describir los de autoridades de modo más específico que SKOS. MADS se diseñó para complementar y suplir las carencias de SKOS como sistema generalista de descripción de vocabularios y para ello se ha establecido un mapeo completo entre ambos esquemas que permite asegurar la interoperabilidad de los recursos descritos por ambos estándares, definiéndose como una estructura jerárquica en la que SKOS es la clase y MADS la subclase.

4.4.1 ELEMENTOS BÁSICOS DE MADS

Los principales elementos de la ontología MADS son: la autoridad, los elementos de autoridad no autorizados y las etiquetas de autoridad.

Las descripciones en MADS son declaraciones sobre esos elementos; las etiquetas representan léxicamente cada tipo de elemento de autoridad.

Las principales clases son:

1. *"madsrdf:Authority"*, para representar los conceptos de autoridad.
2. *"madsrdf:DeprecatedAuthority"* que se refiere a una forma no autorizada de autoridad.
3. *"madsrdf:Variant"*, que representa las formas alternativas de una etiqueta de autoridad y que se constituye como un nodo en blanco que se vincula con subclases como *"madsrdf:hasVariant"* que identifican más específicamente el recurso de autoridad.
4. *"madsrdf:MADSType"* es la clase principal de los elementos simples y compuestos.
5. Las clases *madsrdf:MADSScheme* y *madsrdf:MADSCollection*, se utilizan para la descripción de KOS, descripciones de autoridad o recursos respectivamente. Su papel es similar a las respectivas super clases de SKOS, *skos:ConceptScheme* y *skos:Collection*.

Las relaciones en MADS/RDF, siguen paralelamente a las de SKOS. Se describen tanto las relaciones internas como las de mapeo:

1. *madsrdf:hasRelatedAuthority* es equivalente a *skos:semanticRelation* y se comporta como una superclase para el resto de relaciones:
 - a. *madsrdf:hasBroaderAuthority* y *madsrdf:hasNarrowerAuthority* para representar los conceptos generales y específicos de autoridad.

- b. *madsrdf:hasBroaderExternalAuthority*,
madsrdf:hasNarrowerExternalAuthority se comporta como *skos:broadMatch* y *skos:narrowMatch*
- c. *madsrdf:hasCloseExternalAuthority* y
madsrdf:hasExactExternalAuthority, tal como *skos:closeMatch* o *skos:exactMatch*
- d. *madsrdf:hasReciprocalAuthority* y
madsrdf:hasReciprocalExternalAuthority, para establecer relaciones reciprocas y equivalentes entre recursos de autoridad de un mismo esquema o de diferente esquema o vocabulario.

4.4.2 GESTIÓN DE LA PRECOORDINACIÓN CON MADS

MADS no describe las reglas de uso de subencabezamientos en las listas de encabezamientos de materia, pero a través de la gestión de las etiquetas puede concatenarlas como si de un sistema precoordinado se tratara. La gestión de tipos de conceptos se efectúa partiendo de la clase *madsrdf:MADSType*, la cual tiene como subclases *madsrdf:ComplexType* y *madsrdf:SimpleType*. Cuando se definen los elementos de autoridad, etiquetas alternativas de autoridad y las formas no autorizadas, la descripción se introduce bien con un tipo complejo de concepto o uno simple.

El tipo complejo une etiquetas desde por ejemplo, dos descripciones de autoridad, si hablamos de materias utilizaremos la subclase del tipo complejo *madsrdf:ComplexSubject* que concatena dos o más tópicos (*madsrdf:SimpleType*) constituyéndose el encabezamiento precoordinado (este sistema deriva de la descripción de elementos simples y compuestos de XML Schema, de hecho la propiedad *madsrdf:componentList*, se comporta de igual manera de las listas enumeradas de XML Schema).

Para modelar la expresión de la pre coordinación en MADS, según el modelo de la Library of Congress, debemos crear un espacio de nombres para los subencabezamientos y generar dos registros, el propio de la subdivisión (una o varias) y el registro de la expresión pre coordinada. En el caso que nos ocupa vamos a describir la materia “Guerra de Cuba -- 1885-1898”. Se expresa en primer lugar el registro del subencabezamiento:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://www.upm.es/biblioteca/kos/sh/subenc/XX90410>
  madsrdf:authoritativeLabel "1885-1898" ;
  madsrdf:elementList
    ([ madsrdf:elementValue "1885-1898" ;
      a madsrdf:TemporalElement ] ) ;

  madsrdf:isMemberOfMADSScheme
<http://www.upm.es/biblioteca/kos/sh/subenc> ;
  a madsrdf:Authority, madsrdf:Temporal .
```

Como se puede observar, MADS hace una utilización intensiva de las agrupaciones de conceptos, siendo habitual agrupar los diferentes elementos del concepto precoordinado en una colección MADS, tal y como se ve en la expresión del registro del encabezamiento compuesto (Library of Congress, 2012a; Library of Congress, 2012b).

```
.

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://www.upm.es/biblioteca/kos/sh/XX5454>
  a madsrdf:Topic ;
  skos:preflabel "Guerra de Cuba" .

<http://www.upm.es/biblioteca/kos/sh/XX90410>
  madsrdf:authoritativeLabel "Guerra de Cuba -- 1895-1898" ;
  madsrdf:componentList
    (<http://www.upm.es/biblioteca/kos/sh/XX5454>
    <http://www.upm.es/biblioteca/kos/sh/subenc/XX5265>) ;

  madsrdf:isMemberOfMADSScheme <http://www.upm.es/biblioteca/kos/sh/>,
  <http://www.upm.es/biblioteca/kos/sh/subenc/>;

  a madsrdf:Authority, madsrdf:ComplexSubject .

<http://www.upm.es/biblioteca/kos/sh/subenc/XX5265>
  a madsrdf:Temporal ;
  skos:preflabel "1885-1898"
```


MADS descompone en partes el encabezamiento y lo trata como *ComplexSubject*, ensamblando varios *Simpletype* con valor de autoridad. En el caso del ejemplo el subencabezamiento es de tipo específico, pues sólo se puede combinar con la entrada de autoridad “Guerra de Cuba”.

4.5 VOCABULARIOS DE MATERIAS. ANÁLISIS DE CASOS DE USO

Las listas de encabezamientos de materia representan uno de los productos más específicamente bibliotecarios. Como vocabulario controlado presenta la indudable ventaja de la calidad de los datos que las componen y de su eficacia como punto de acceso a los recursos bibliográficos. En cambio, este tipo de vocabularios, no dispone de la “popularidad” de otros productos para la organización del conocimiento, como los tesauros. Las listas de encabezamientos de materia no tienen un estándar internacional que las regule, a lo sumo existen reglas y pautas con cierto carácter local y de publicación relativamente reciente. Ciertamente existen manuales de desarrollo en los principales sistemas, pero sin ninguna aspiración de universalidad y sin la fuerza cohesionadora de una norma internacional. En el contexto Linked Data y su literatura, no hay demasiadas pautas para la publicación en concreto de estos vocabularios, aunque los vocabularios de materias, curiosamente han sido pioneros en cuanto a su aparición en la Web de datos. Se requiere pues un breve repaso de las mejores pautas y buenas prácticas de la publicación de las LEM en Linked Data en los principales sistemas bibliotecarios.

4.5.1 LIBRARY OF CONGRESS SUBJECT HEADINGS

La lista de encabezamientos de la Library of Congress (LCSH) es el más importante vocabulario de materia en inglés a nivel mundial, dada su alta calidad técnica y su amplio ámbito de uso. Su versión en Linked Data persigue y fomenta la reutilización de este vocabulario de manera totalmente abierta, presentando un formato legible por máquinas y humanos. Los LCSH están publicados en multitud de formatos, entre otros MADS/RDF y SKOS. (Library of Congress, 2014b)

La Library of Congress Subject Headings reemplaza la típica organización de los sistemas de encabezamientos de materia, por una estructura en tesoro, proceso que se intentará replicar aquí bajo las directrices de la ISO 25964. Su estructura concisa comprende los siguientes puntos:

1. Relaciones de jerarquía: BT, NT. Para cada relación BT se generan una o varias NT.
2. Relaciones asociativas, RT, cuyos tipos de especificación fundamentales son:
 - Disciplina objeto de estudio
 - Personas actividades
 - Significados que se superponen

3. Relaciones de equivalencia: las referencias USE se utilizan desde términos no preferidos a encabezamientos de materia autorizados. Los no autorizados no se pueden utilizar para indizar.
4. Referencias generales: vinculan un encabezamiento a un grupo de encabezamientos.
5. Notas de alcance: especifican el alcance y rango de aplicación y para establecer distinciones entre encabezamientos relacionados (Library of Congress, 2014b).

Las LSCH utilizan cuatro sistemas de precoordinación (subdivisión, inversión, calificación y construcción de la frase). Cada método indica una relación semántica específica entre las materias coordinadas.

En primer lugar y tal como exige el modelo de datos de SKOS, se necesitaba identificar cada concepto con un IRI. Dado que los requerimientos semánticos más elementales indican la necesidad de aplicar un IRI permanente, se introdujo como código identificador del concepto el número de cada materia en la Library of Congress Control Number. Este número es estable y normalizado y se incluyó como parte de la IRI, dándole un elevado nivel de estabilidad.



Figura 19 Utilización Library of Congress Classification Number. Fuente: elaboración propia

El mapeo de encabezamientos de materia autorizados y no autorizados se hizo fácilmente a través de la conversión en `skos:prefLabel` y `skos:altLabel`, aunque parte de la riqueza descriptiva de MARC se perdía por la inexpresividad, o mejor dicho, generalidad semántica de SKOS, por ejemplo no se especificaban los tipos de encabezamientos: cronológicos, temáticos, geográficos, de género o forma. Otra de las dificultades se basaba en la incapacidad de SKOS de generar encabezamientos precoordinados, a lo sumo establecía entradas simples y entradas compuestas. También surgieron problemas en la adaptación de vocabularios multilingües, al existir conceptos expresados en varios idiomas, no aparecía indicación alguna de cual es preferente respecto al otro. Respecto a las relaciones semánticas era preciso ajustar las que se establecen entre conceptos generales y específicos, pues SKOS requiere la asignación en ambos conceptos. Para la descripción de metadatos fuera del alcance de SKOS se recurrió a vocabularios externos como Dublin Core. Respecto a la implementación, apuntar que el proceso se basó en una secuencia consistente en la precisión de la estructura conceptual, generación de IRIs y integración de ambas partes en los registros de encabezamientos.

En la actualidad la colección de vocabularios en Linked Data de la Library of Congress es la más amplia del mundo y sigue actualizándose, aportando hoy un verdadero vocabulario mundial multilingüe en cuatro idiomas: inglés, francés, alemán e italiano. Su proceso de adaptación tecnológica continúa, ofreciendo serializaciones en los formatos más actuales, aunque sigue sin facilitar un servicio de recuperación como SPARQL, al menos de modo directo, aunque si permite la descarga del dataset y su posterior procesado por un servicio SPARQL endpoint de modo local (Library of Congress, 2014b; Summers, Isaac, Redding, & Krech, 2008).

4.5.2 RAMEAU

El Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU, <http://www.cs.vu.nl/STITCH/rameau/>), es el listado de encabezamientos de materia de la Biblioteca Nacional de Francia, expresado mediante SKOS. La representación se efectúa a través de un servicio todavía en fase beta, STITCH (Semantic Interoperability to Access Cultural Heritage) y conforma una estructura multilingüe al vincularse con las LCSH y los encabezamientos de materia de la Biblioteca Nacional de Alemania (Schlagwortnormdatei, SWD). El servicio suministra los datos vinculados en lenguajes RDF o HTML; como novedad, ofrece marcado semántico embebido en RDFa.

Su sistema de procesado es más sencillo que el de la Library of Congress, por ejemplo emplea el mismo tratamiento para los encabezamientos precoordinaados que para el resto, expresándolos simplemente en `skos:prefLabel`. Como complementos a la descripción utiliza lenguajes como OWL o Dublin Core y se ha gestionado su propia ontología “Onto-FRBNF”. Técnicamente utiliza el sistema “ark” para los IRIs y serialización en RDF/XML. Mapea con el “namespace” de Dewey, generando un vínculo por cada elemento del encabezamiento.

```
<skos:prefLabel xml:lang="fr">Guerres napoléoniennes (1800-1815) -- Projet  
d'invasion de l'Angleterre (1793-1805)</skos:prefLabel>
```

(Bibliothèque nationale de France, 2014a).

4.5.3 SCHLAGWORT NORMDATEI (SWD)

La Biblioteca Nacional de Alemania también dispone de un servicio Linked Data que ofrece entre otros registros de autoridad, sus encabezamientos de materia (“*Schlagwortnormdatei*”, SWD). Su aspiración es exponer todo su material bibliográfico en el nuevo formato y permitir su acceso a través de canales de acceso ya establecidos como OAI o SRU (Search Retrieve via URL, protocolo estándar XML para la generación de ecuaciones de búsqueda). Propone un protagonismo dominante de las bibliotecas y sus datos, interactuando, no sólo con los agentes habituales, sino

también con proveedores de servicios comerciales, motores de búsqueda, organizaciones de investigación, etc.

Su propuesta para la reutilización de los datos es total, como lo demuestra el tipo de licencia del sistema de datos vinculados del SWD es la Creative Commons Zero ("*Public Domain*"). Los datos vinculados de autoridades de la SWD se presentan en el GND (repositorio de autoridades), representando a personas, entidades, materias, etc., y se ofrecen como sistema de referencia en cuanto a la descripción de recursos en el mundo del patrimonio cultural y otras organizaciones científicas y culturales. El vocabulario preferente es RDF/SKOS, pero se utilizan ontologías propias como la GND Ontology para una mayor granularidad en la descripción de las autoridades (Deutsche Nationalbibliothek, 2014a).

4.5.4 NUEVO SOGGIETTARIO

El NS es una herramienta para la indización con materias en el ámbito del Sistema Bibliotecario Nacional de Italia. Técnicamente puede ser aplicado tanto a sistemas precoordinaados como a post coordinados. El NS se nutre de los encabezamientos de materia del sistema tradicional, de las adiciones de los indizadores y de los que se deducen a través de las redes de relaciones semánticas. Desde un punto de vista de los datos vinculados, el NS se conecta para su enriquecimiento semántico, con Wikipedia y otras bases de datos, además establece equivalencias con encabezamientos de la Library of Congress. Sus desarrollos hasta la actualidad abarcan la publicación en Linked Data con SKOS, pero no sólo eso, se pretenden establecer procesos estables y automáticos para la indización, lo que lógicamente disminuirá las cargas de catalogación.

Al igual que la Library of Congress, la Biblioteca de Florencia ha tenido problemas con los ajustes entre la estructura de encabezamientos y el alcance de la semántica de SKOS. También se ha tenido reconsiderar la relación entre encabezamientos nuevos y en desuso, del mismo modo y respecto a los encabezamientos de materia complejos y sus remisiones a descriptores y no descriptores (USE/USADO POR), se han tenido dificultades para establecer relaciones entre multitérminos y términos simples, para lo que se trabaja en una posible solución a través de la extensión SKOS-XL. La interoperabilidad con otros vocabularios se establece bajo dos premisas: creación en los registros de materia de un campo "fuente", si el vocabulario objetivo está modelado en SKOS se utilizan las propiedades `skos:closeMatch` para enriquecimiento; si no está disponible, se utiliza el enlace en el campo "fuente" para acceder al vocabulario referenciado; el segundo sistema, también mediante `skos:closeMatch`, se establece un campo para reflejar las equivalencias entre vocabularios. Ofrece serializaciones en RDF/XML y cuenta también con su propia ontología de apoyo.

4.6 VINCULACIÓN ENTRE VOCABULARIOS

En otras partes de este trabajo se ha hablado de la posibilidad de reutilizar varios vocabularios para obtener descripciones más precisas. En el contexto de los vocabularios determinados para KOS este mismo procedimiento permite aumentar la riqueza en la descripción, sumando la riqueza expresiva de esquemas de muy diferente concepción.

Como se dijo ha dicho, SKOS es un estándar de modelado generalista y por ello se requiere en ocasiones la utilización combinada de diferentes vocabularios previamente alineados y el mapeo entre los diferentes esquemas y conceptos. La riqueza que esa mezcla supone, aumenta cuando ajustamos el valor semántico de dichos vocabularios mediante propiedades que permiten enlazar los conceptos. Esto propicia que tanto los sujetos como los objetos de las tripletas puedan ser usadas varias veces vinculados con propiedades, en diferentes datasets estableciendo así los cimientos de la interoperabilidad entre dichos esquemas.

Dunsire propone una técnica de mapeo global entre los vocabularios más comunes, utilizando sus propiedades expresadas en RDF y asociándolas con otras propiedades o clases mediante reglas de inferencia, lo que en su opinión permite una “comunicación” entre las diferentes propiedades y clases de los vocabularios. Las propiedades interrelacionadas (por ejemplo mediante *rdfs:subpropertyOf* o *skos:broader*) se constituyen como una nueva declaración RDF cuyos elementos están en diferentes vocabularios y todo el conjunto de mapeos podría considerarse como una ontología global (Dunsire et al., 2012).

Por ejemplo, podemos asociar propiedades de diferentes esquemas expresando jerarquía entre ellas que permitiendo una mayor especificación de la descripción. La propiedad ISBD “*hasAdditionToPlaceOfPublication*” que relaciona un recurso con el país o Estado de la publicación, puede definirse como subpropiedad de la propiedad RDA “*placeOfPublication*”, que hace referencia al lugar concreto con la publicación.

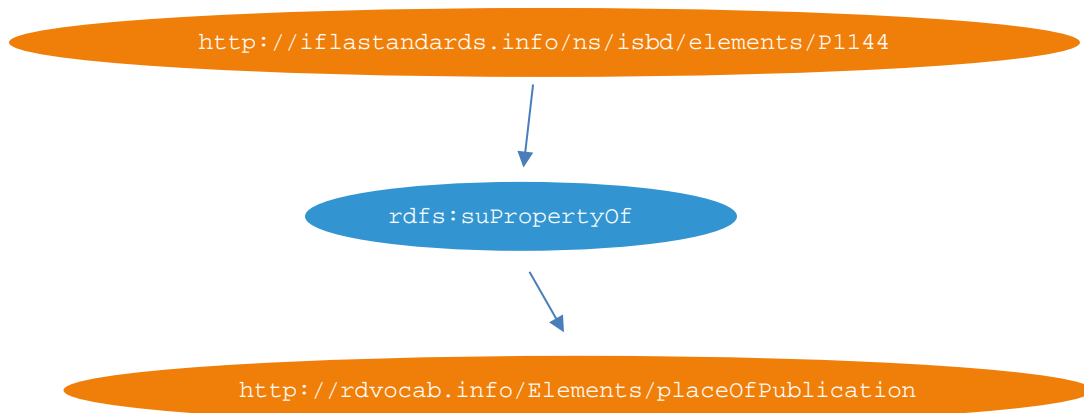


Figura 20 Relación de mapeo entre propiedades de diferentes vocabularios. Fuente: elaboración propia.

A efectos de los vocabularios de valores, el alineamiento supone el establecimiento de relaciones entre entidades similares en diferentes esquemas, lo que permite una identificación más eficaz de las relaciones entre los componentes de esos esquemas. Por ejemplo y tal y como propone Miklos (2012), podemos alinear propiedades de SKOS Y OWL que permitan la transformación de esquemas de conceptos en ontologías formales:

- skos:Concept se puede transformar en owl:Class
- skos:prefLabel se puede transformar en rdfs:label
- skos:broader se puede transformar en rdfs:subClassOf
- skos:definition se puede transformar en rdfs:comment

Los esquemas OWL, SKOS y RDFS proporcionan las propiedades necesarias para realizar alineaciones entre vocabularios de valores y/o esquemas de metadatos. El mapeo de las propiedades pueden también ser generadas de modo automático y dinámico, permitiendo la generación de vínculos entre elementos de diferentes esquemas según las necesidades de descripción o modelado (Dunsire & Willer, 2013).

Tabla 6 Principales propiedades para el mapeo de vocabularios. Fuente: elaboración propia.

| | |
|---------------------------|---|
| owl:sameAs | Los elementos relacionados son iguales y tienen el mismo valor. Pueden utilizarse indistintamente. |
| rdfs:subClassOf | El objeto de un esquema tiene un alcance semántico mayor que el sujeto de otro esquema. |
| rdfs:subPropertyOf | |
| skos:exactMatch | Los conceptos representados en los sujetos y objetos de los diferentes esquemas tienen el mismo significado. |
| skos:closeMatch | Los conceptos representados en los sujetos y objetos de los diferentes esquema tienen un significado similar y pueden utilizarse indistintamente en las aplicaciones. |
| skos:broadMatch | El concepto que es objeto de la tripleta tiene un alcance semántico mayor que el sujeto. |
| skos:narrowerMatch | El concepto que es objeto de la tripleta tiene un alcance semántico menor que el sujeto. |
| skos:relatedmatch | Los conceptos de los sujetos y objetos de los esquemas tienen significados relacionados. |

5 VOCABULARIO DE MATERIAS DE LA BIBLIOTECA DE LA UNIVERSIDAD POLITÉCNICA DE MADRID (BUPM) BAJO LAS DIRECTRICES Y RECOMENDACIONES DE LINKED OPEN DATA

5.1 PLANIFICACIÓN GENERAL

En este capítulo se pretende construir un procedimiento viable para la publicación del vocabulario de materias de la Biblioteca de la Universidad Politécnica de Madrid (BUPM) bajo las directrices y recomendaciones de Linked Open Data. Se trata de un prototipo de cerca de 1000 encabezamientos, cuyo contenido refleje un marco conceptual básico que pueda servir de base para futuros desarrollos o implementaciones.

Para su realización se han analizado los conceptos fundamentales para la transformación de un vocabulario de conceptos temáticos, en una estructura abierta y conformada como un conjunto de datos reutilizables y combinables con otros vocabularios de similar finalidad y en diferentes idiomas. Se ha pretendido ofrecer y conformar una plataforma multilingüe que reúna las peculiaridades semánticas derivadas de las diferencias culturales de los conceptos equivalentes y relacionados en los diferentes vocabularios involucrados.

Se analiza, en los siguientes epígrafes, aquellos puntos centrales de la planificación previa del proyecto: fijar unos objetivos alcanzables, evaluar las necesidades mostradas por la comunidad de usuarios del vocabulario, definir las razones que justifican la realización de este proyecto y el inventario y elección de los recursos disponibles.

5.1.1 ANÁLISIS DE LA COMUNIDAD DE USUARIOS

Las materias, en el contexto de trabajo y uso de la BUPM, son utilizadas por la comunidad universitaria fundamentalmente para la búsqueda de información y por los profesionales bibliotecarios también para la indexación de las colecciones y recursos que integran los fondos de la biblioteca.

El objetivo de cualquier LEM para conseguir un nivel de calidad adecuado es captar el punto de vista de quién las utilizan, compatibilizarlo con las mejores prácticas a la hora de generar el vocabulario y conseguir un grado de utilización en los sistemas de búsqueda aceptable, cuestión esta última de muy difícil consecución.

La migración a Linked Open Data no pretende cambiar la configuración actual del vocabulario de materias, pero si podría ser una oportunidad para mejorar los aspectos antes señalados, procurando una mayor eficiencia en la ingesta de nuevos conceptos y una mayor visibilidad del vocabulario en general, cuestiones que aumentarían probablemente su uso. No se trata de abordar aquí un nuevo procedimiento de actualización, pero si, al menos, se pueden sugerir algunas recomendaciones que mejoren el ajuste entre el vocabulario y la comunidad a la que sirve.

En primer lugar, se debería diferenciar entre los diferentes grupos integrantes de la comunidad para atender sus específicas necesidades y el nivel de materias que les serían más útiles: usuarios investigadores, alumnos en su fase de formación, áreas temáticas comunes, nivel de granularidad necesario en los conceptos. Por otro lado sería beneficioso establecer sistemas de recogida de información sobre la utilización de materias, como los análisis de los ficheros *log* que generan los catálogos y otras sistemas de recuperación como metabúscadores o bases de datos. También sería conveniente establecer contactos directos con representantes de la comunidad que aporten su experiencia durante la búsqueda o la indización; para ello sería conveniente establecer un sistema de entrevistas regulares y unos protocolos sostenidos para la participación de expertos, especialmente entre el personal docente. Estas breves recomendaciones ayudarían, al enriquecimiento y aumento de la calidad de los vocabularios de materias, cuestión de vital importancia para la Web de datos (International Organization for Standardization (ISO), 2011; Leroi, Holland, & Cagnet-Dédale, 2011).

5.1.2 ANÁLISIS DE RECURSOS

Una estructura consistente para la creación o modificación de un vocabulario de materias debería contener los recursos humanos y materiales necesarios para la organización y desarrollo del proyecto en todos sus aspectos. En un supuesto ideal, la estructura de los recursos humanos debería definir la responsabilidad del editor del vocabulario y de su equipo en las diferentes etapas de desarrollo, contando con un plan de control de calidad que asegure los buenos resultados. Estos recursos deben completarse con la inclusión de usuarios que participen en la evaluación de la calidad en las diferentes fases.

Para las tareas de definición conceptual y del sistema semántico de relaciones se requiere personal experto en las áreas del conocimiento afectadas, para la gestión del multilingüismo se requieren profesionales con conocimientos avanzados de los idiomas involucrados, también personal con competencia en la gestión de sistemas de organización del conocimiento y utilización de herramientas semánticas, y finalmente, personal de sistemas para la actuación en la implementación computerizada.

Los recursos materiales y humanos para el desarrollo de un proyecto de esta índole se han ajustado al específico contexto instructivo en el que se encuadra dentro de la Universidad Carlos III de Madrid:

1. Utilización de herramientas de software libre o bajo licencias de uso de la Universidad.
2. Acceso a fuentes de información de pago mediante las pasarelas de acceso.
3. Recursos web de acceso libre.
4. Apoyo docente y del personal bibliotecario de la Universidad.
5. Y como herramienta básica, el gestor de tesauros, en la versión de licencia académica.

Este último punto es de gran importancia, pues acometer un proyecto de esta índole de modo exclusivamente manual no es operativo ni eficiente, sobre todo respecto a la calidad esperada del producto.

5.1.2.1 Elección del software para la gestión semántica del tesoro de materias

La elección de un sistema de gestión de tesauros no es una tarea fácil, probablemente la mejor opción sería contar con personal de sistemas y expertos en programación que pudieran ofrecer “ex novo” una plataforma adaptada a las necesidades específicas de cada proyecto. Las consideraciones generales a evaluar podrían ser, sin ánimo de ser exhaustivos:

1. El rango de funcionalidad.
2. Las capacidades que ofrece desde el punto de vista del vocabulario (actualización, reducción de errores, facilidades de importación y exportación, etc.).
3. La propuesta de almacenamiento y su fiabilidad.
4. El grado de complejidad de su manejo, instalación o de su aprendizaje.
5. La modularidad y flexibilidad como capacidad de adaptación a las necesidades actuales y futuras.
6. Las capacidades semánticas de representación.
7. Las posibilidades de publicación.
8. La herramientas para la preservación de los datos.

(Martínez-González & Alvite Díez, 2014; Morshed & Rittaban, 2012).

Por cuestiones de economía de los procesos se han evaluado las herramientas para la gestión de tesauros TEMATRES y PoolParty. Respecto a la primera reconocer que es una herramienta útil pues permite la gestión de un tesoro a nivel básico de un modo sencillo, contando además con la ventaja de su gratuidad. Como inconveniente, su mayor nivel de complejidad en la instalación pues requiere el montaje de un servidor web y la configuración de algunos parámetros en PHP. Teniendo en cuenta estos aspectos principales, se decidió no utilizar dicha aplicación por carecer en su versión actual de ningún soporte para mapeos de vocabularios, cuestión de importancia nuclear en este proyecto, además de otras cuestiones como la falta de gestión de metadatos,

menores capacidades de trabajo con corpus definidos, inferiores capacidades de exportación y de importación y ausencia de sistemas de comprobación de la calidad.

Teniendo en cuenta los criterios anteriores, el producto elegido es PoolParty Thesaurus Manager, de Semantic Web Company, herramienta basada en una plataforma web y que permite, en su versión académica (la utilizada aquí), la gestión de metadatos, la publicación de los vocabularios en Linked Data en diferentes formatos, testeo de vocabularios y extracción terminológica mediante la minería de datos, (Semantic Web Company GmbH, 2013).

Un análisis más profundo se ha realizado contrastando las recomendaciones para aplicaciones gestoras de tesauros reguladas por la norma ISO 25964 (epígrafe 14) y las especificaciones del estándar SKOS, determinándose las siguientes conclusiones (International Organization for Standardization (ISO), 2011; Isaac & Summers, 2005; Semantic Web Company GmbH, 2013):

1. Restricciones de tamaño o de tipo de caracteres:

- a. No limita la cantidad de términos, longitud de cadenas textuales o profundidad en los niveles de jerarquías.
- b. Se refiere en las especificaciones la capacidad de presentar el juego de caracteres ISO/IEC 10646.
- c. No tiene limitación para incluir mayúsculas o minúsculas.
- d. Su rango de idiomas se limita a los más comunes.

2. Relaciones entre los términos:

- a. No permite conceptos duplicados en el mismo lenguaje, dentro del mismo *ConceptScheme* (aviso en *Report Quality*).
- b. Admite las relaciones básicas de jerarquía, equivalencia y asociativas.
- c. Es coherente en las relaciones de reciprocidad entre los conceptos en cualquiera de sus tipos, aunque permite la introducción de entradas erróneas avisando del conflicto a través de la ejecución manual del *Report Quality*.
- d. Las eliminaciones de conceptos se propagan correctamente.
- e. No permite la distinción de diferentes tipos de transitividad en las jerarquías o asociaciones.
- f. Permite la inserción de relaciones asociativas entre términos generales y específicos, avisando del error mediante *Report Quality*.
- g. No permite relaciones de términos no preferentes.
- h. No permite la relación entre el término preferente y su concepto.
- i. No permite más de un término preferente por idioma.

3. Anotaciones

- a. Permite únicamente notas de alcance, definiciones y ejemplos.

4. Categorías, notaciones y etiquetas de nodo

- a. Contempla todas las posibilidades de agrupación de conceptos que permite tanto la norma ISO 25964 como el estándar SKOS.
- b. Permite la generación de códigos identificativos asociados al IRI.
- c. No permite la utilización de etiquetas de nodo.

5. Multilingüismo

- a. Cabe la posibilidad de introducir etiquetas en diferentes idiomas para el mismo concepto.

6. Importación y exportación

- a. Las características de variedad de formatos de salida y entrada están bien soportadas en PoolParty. Existe una amplia variedad de formatos de serialización tanto de la familia RDF como de otros estándares de codificación de vocabularios.
- b. Cabe la importación desde Excell.
- c. Se pueden exportar los datos sin modelado.
- d. Los tipos de datos que se pueden exportar son: los conceptos, el flujo de trabajo, el historial de trabajo, las listas de recuperación de SPARQL, y los ficheros VOID y ADMS.
- e. Se contempla la exportación del proyecto al completo, pero no se hace referencia a la posibilidad de exportar la estructura del tesoro a un formato no propietario, como las notas o relaciones de todo tipo.
- f. No están disponibles filtros para la exportación.

7. Edición

- a. Las operaciones de edición o borrado por lotes no son posibles.
- b. Se puede navegar por los enlaces visualizados en jerarquías u otro tipo de relaciones.
- c. Tiene capacidad para mover tramos de jerarquías completos.

- d. Existen varias opciones de visionado de los datos, presentando una aplicación visual de regular calidad.
- e. Existe ayuda contextual a la introducción texto.
- f. Es posible la coedición.

8. Seguridad

- a. El sistema permite el autoguardado automático.
- b. El software tiene control de acceso por autenticación.

(International Organization for Standardization (ISO), 2011)

Otras ventajas que podemos afirmar a parte de las verificadas con la norma son: la facilidad para el manejo de la interfaz y para el aprendizaje de los procesos más habituales (aunque para su manejo con soltura requiere una cierta comprensión del estándar SKOS); el establecimiento de mapeos cuasi automáticos con vocabularios de referencia como Geonames, Dbpedia o LCSH; la posibilidad de contar una interfaz de recuperación SPARQL en la publicación y la generación de una interfaz en modo “wiki” que permite una adecuada representación en la Web.

En cambio, algunas características importantes se echan en falta en la plataforma, probablemente porque nos encontramos ante una solución orientada a la publicación de tesauros conceptuales y no listas de encabezamientos de materia:

1. Es posible la utilización de colecciones SKOS, pero el marcado semántico que se establece se basa en ontologías propietarias.
2. No se pueden hacer declaraciones de transitividad en las jerarquías.
3. No se pueden incluir otros vocabularios o extensiones para completar la descripción, como por ejemplo MADS o SKOS-XL.
4. No existen posibilidades de mapeo libres, a los efectos del proyecto se puede realizar mapeo automático únicamente con la LCSH, por lo que la inclusión de mapeos con el resto de vocabularios se efectuará de modo selectivo y manual.

5.2 IMPLEMENTACIÓN DEL VOCABULARIO

El proyecto que nos ocupa ha tratado de transformar un vocabulario de materias en un tesauro de acuerdo a normas estandarizadas. No se trata de generar un vocabulario *ex novo*, más bien una actualización hacia formatos conceptuales adecuados para su representación en la Web y descritos y publicados bajo técnicas Linked Data.

En los siguientes epígrafes se contiene la fase nuclear del proyecto, abarcando etapas muy diversas, desde la obtención de los *raw data*, hasta la publicación de los datos en la Web, pasando por todas las fases intermedias, como los análisis, la definición del dominio, el análisis y ajuste a las normas, el modelo de datos, control de vocabulario y relaciones, mapeos, asignación de IRLs y la “skosificación”.

5.2.1 OBTENCIÓN DE LOS DATOS PARA LA CONSTRUCCIÓN DEL VOCABULARIO

Los datos fueron obtenidos previa petición al Servicio de Coordinación de las Bibliotecas UPM, a la que se unió una consulta sobre las cuestiones relativas a la de propiedad intelectual que incluso en el caso de un proyecto prototipo se podrían substanciar. La petición de los datos fue satisfecha mediante el suministro de un reporte de datos en formato texto con los registros formateados en MARC de autoridades.

En primer lugar, respecto a las cuestiones relativas a la propiedad intelectual, se analizan las opciones del proceso que pueden afectar a posibles derechos de autor sobre el listado de materias. Las listas de encabezamientos de materia de la BUPM han sido formadas e integradas a través de los años por las aportaciones del personal técnico de los servicios, a los cuales no se les reconoce derecho alguno por la contribución aportada con independencia del trabajo de fundamentación requerido. Se reconoce en cambio el derecho de propiedad intelectual sobre los listados de materias a la Universidad Politécnica de Madrid, como ente que ha concretado el producto intelectual de varios agentes hasta la configuración del sistema actual de materias. Como única restricción al uso se prohíbe la utilización comercial de los datos suministrados, tampoco se contempla la asignación previa de una licencia.

5.2.2 ANÁLISIS DE LOS REGISTROS DE MATERIA

Los registros entregados se analizan para verificar la estructura y el contenido:

1. Campo 1º: información de clasificación de la materia según las diferentes áreas de actividad de las Escuelas y Facultades, estableciéndose cinco grupos temáticos específicos y uno general.
2. Campo 2º: etiqueta 001, identificador unívoco del registro de materia. Se trata de un identificador que contiene dos caracteres fijos (XX) y alternativamente cinco o seis dígitos (e.g. XX252589)
3. Campo 3º: etiqueta 150, encabezamiento de materia aceptado.
4. Campo 4º: etiqueta 450, campo de referencia de envío (UF) para las formas no autorizadas.

5. Campo 5º: etiqueta 550, campo de referencia de envío (RT) para los encabezamientos relacionados con el encabezamiento del campo 150.
6. Campo 6º: etiqueta 670, fuente consultada para la creación del encabezamiento y código de subcampo \$b, para la inclusión de la información localizada, es decir el encabezamiento de materia del esquema remoto.
7. Campo 7º: etiqueta 680, para la fuente general de información al usuario.

No se han incluido en el reporte los siguientes campos MARC presentes en el sistema de gestión de la biblioteca:

1. Etiqueta 005: Fecha y hora del último cambio del registro de autoridad.
2. Etiqueta 065: Número de Clasificación.
3. Etiqueta 080: Número de la Clasificación Decimal Universal.
4. Código de subcampo \$y de la etiqueta 150, subdivisión cronológica (subencabezamiento cronológico)
5. Etiqueta 750: Contiene un término temático equivalente al encabezamiento incluido en el Campo 150. Con este campo se establece un vínculo entre dos encabezamientos de materia que pertenecen a un mismo o a diferentes sistemas, tesauros o archivos de autoridades.

No existe uniformidad en la de descripción de los registros suministrados, siendo la única etiqueta común a todos ellos la 150.

5.2.2.1 Control de errores por campos

Se procede a la limpieza de errores evidentes en la identificación de los campos, necesaria para definir los campos semánticos disponibles. Se trata de asignar un mismo código de área a cada registro, corrigiendo los errores percibidos y filtrando los conceptos por dicho código. Por ejemplo la configuración estándar del Campo 1º incluye las siglas de la UPM y el identificador de área (A1, A2) separados por un guion; se tiene esta forma como normalizada y se corrigen todas las demás (e.g. UPM-A1, forma normalizada; UPM A1 forma incorrecta, UPM-A-1, forma incorrecta).

5.2.2.2 Filtrado de descriptores de materia

Corregidos los errores en el Campo 1º se efectúan sucesivos filtros para separar los descriptores de materia según las áreas de actividad codificadas en el registro. Tras aislar las diferentes áreas, se filtran los conceptos alfabéticamente para detectar relaciones. Todas estas agrupaciones se ordenan sistemáticamente y se estudia el campo semántico recuperado, intentando descubrir las

relaciones semánticas subyacentes: conceptos cabecera, jerarquías y subjerarquías, relaciones implícitas, etc. Para ello se consulta en cada concepto elegido el registro en el propio SIGB, que contiene en ocasiones relaciones de equivalencia y asociativas. También se analizan los grupos de descriptores intentando detectar conceptos y sus etiquetas preferentes.

5.2.3 DEFINICIÓN DEL DOMINIO DEL VOCABULARIO Y DE SU ESTRUCTURA GENERAL

La Universidad Politécnica de Madrid desarrolla su actividad en las diferentes disciplinas que integran los estudios de ingeniería que en sus centros se imparten. La BUPM se estructura fundamentalmente a través de bibliotecas de centro que gestionan la asignación de materias dentro de una cobertura temática específica y asociada al campo de conocimiento particular de cada titulación. Así, los dominios en los que se desarrollará este vocabulario se corresponden con las siguientes áreas de actividad:

Área 1: Arquitectura e Ingeniería Civil.

Área 2: Ingeniería Industrial y Minera.

Área 3: Ingeniería Agroforestal.

Área 4: Ingeniería Aeronaval.

Área 5: Ingeniería de las T.I.C. (Telecomunicaciones e Informática)

Área 6: General, materias comunes a todos los campos, especialmente: Matemáticas, Física, Legislación, Lengua Inglesa, Lengua Española, Lingüística.

Esta configuración permite el diseño de una estructura múltiple con diferentes esquemas o microtesauros que posibilitan una más adecuada sistemática de las materias correspondientes a cada área, procurando establecer entre ellas relaciones que enriquezcan el vocabulario. La recomendación del programa Linked Heritage de la Unión Europea, indica la necesidad de precisar el vocabulario con preferencia a tesauros más pequeños concentrados en el desarrollo de dominios específicos que luego pueden unirse mediante mapeos. Esta solución es más escalable y de más fácil mantenimiento que un gran tesoro, con un complejo sistema de relaciones y es la que en lo posible se ha seguido aquí (Leroi et al., 2011).

5.2.4 ESTABLECIMIENTO DE UN MODELO DE DATOS

El modelo que se representa define las estructuras conceptuales del vocabulario, sus propiedades y atributos, relaciones con otros datos dentro o fuera del vocabulario y las agrupaciones semánticas posibles. Se pretende recoger de modo general todas las posibilidades de información que se produzcan durante su utilización. Se escoge como fundamento teórico, el modelo de datos ISO 25964, adaptado a las necesidades propias de las LEM BUPM conforme a los estándares antes definidos.

El modelo de datos refleja en primer lugar la estructura general del tesoro de materias en su conjunto, relacionado mediante mapeos con el resto de vocabularios de materias escogidos.

La estructura particular refleja la configuración interna del tesoro, partiendo de las macroestructuras o microtesoros que vertebran el dominio del vocabulario. La definición de cada *ConceptScheme* representa la disposición de conceptos generales (*TopConcepts*) y conceptos dependientes de éstos, todos ellos representados léxicamente por etiquetas, que identifican al concepto de modo preferente, las que lo hacen de modo alternativo a efectos de recuperación de la información y las que no aparecen en las interfaces de uso, pero que están contenidas en la estructura léxica como alternativas. También se contempla la agrupación de interés para el contexto del tesoro de conceptos en *arrays* o colecciones.

Todo este sistema se ordena mediante relaciones de supra o subordinación, de equivalencia y asociativas, que se plasman tanto a nivel interno del tesoro como a nivel externo mediante mapeos entre conceptos de los diferentes vocabularios.

Finalmente el sistema de anotaciones se vincula con los conceptos expresando su campo semántico, sus aspectos editoriales, temporales, de procedencia o cuestiones referidas a la sintaxis y morfología de conceptos y términos (International Organization for Standardization (ISO), 2011).

Cabe la posibilidad de controlar la información de los registros normalizando su expresión semántica mediante enlaces a vocabularios de referencia como Geonames, o VIAF (en los casos de utilización de nombres personales en calidad de materia). De este modo los delimitadores geográficos o conceptos personales podrían ser sustituidos por IRIs que conectarán con los contenidos de los vocabularios aludidos. Dicha conexión enriquecería de modo importante el contenido del tesoro de materias. Estas vinculaciones supondrían una evolución en cuanto a usabilidad de los conceptos sobre todo en interfaces de uso humanas.

En el ejemplo anterior se ha vinculado el concepto “Denominación de origen Jerez” en primer lugar y por equivalencia compuesta con los conceptos genéricos *Marks of origin* y Jerez, en LCSH. Para enriquecimiento se añaden vinculaciones a los conceptos “Denominación de origen y Jerez” en DBpedia y para información de localización se introduce el IRI correspondiente en Geonames.


```

<http://academia6.poolparty.biz/Tesauro-Materias-BUPM/580> a skos:Concept ;
    skos:prefLabel "Denominación de origen Jerez"@es ;
    skos:notation "347.774"@es ;
    skos:broadMatch <http://id.loc.gov/authorities/sh85081371#concept> ,
    <http://dbpedia.org/resource/Denominaci%C3%B3n_de_Origen> ,
    <http://dbpedia.org/resource/Jerez_de_la_Frontera_(C%C3%Aldiz)> ;

    skos:relatedMatch <http://id.loc.gov/authorities/sh90004722#concept> ,
    <http://sws.geonames.org/2516326/> .

```

Se establecen unas determinadas reglas editoriales que pretenden servir para dirigir la gestión del vocabulario. El establecimiento de reglas supone siempre documentar aquellos supuestos donde estas son alternativa a los requisitos locales del sistema de bibliotecas BUPM. Unos presupuestos básicos podrían ser los siguientes:

1. Identificación de la información requerida del registro conceptual.
 - a. Obligatorio: concepto, etiqueta preferente, mapeo (si el concepto está en los dos primeros niveles jerárquicos), notación en CDU, nota editorial (si se cambia el estado del concepto, o se añade uno nuevo, o cualquier información pertinente susceptible de incluirse en la información editorial), relaciones jerárquicas obligatorias para *TopConcept*.
 - b. Optativa: etiquetas no preferentes y ocultas, definiciones, relaciones jerárquicas para subsiguientes niveles, relaciones asociativas, relaciones equivalentes, ejemplos y mapeos para otros niveles.
2. Reglas para la elección del término preferente: las especificadas en el vocabulario fuente.
3. Ámbito de inclusión de términos no preferentes: inclusión optativa.
4. Reglas para establecer las posiciones jerárquicas de los términos añadidos: según tipología ISO 25964-1 (10.2)
5. Reglas para la extensión de las líneas de jerarquía: nivel obligatorio para *TopConcept*, para el resto dependiendo del interés semántico y la utilidad en la indización y recuperación.
6. Reglas para establecer relaciones entre términos y otros registros: según ISO 25964-1 (10.3)
7. Formato y sintaxis requerida a incluir en los registros: según ISO 25964-1 Epígrafes 6 y 7.
8. Definición del idioma principal: español con mapeos en inglés, francés, alemán e italiano.
9. Fuentes autorizadas para los diferentes campos del conocimiento. (Enumeración incorporada en el capítulo introductorio, apartado 1.4)

10. Árboles de decisión para escoger entre información contradictoria entre fuentes.

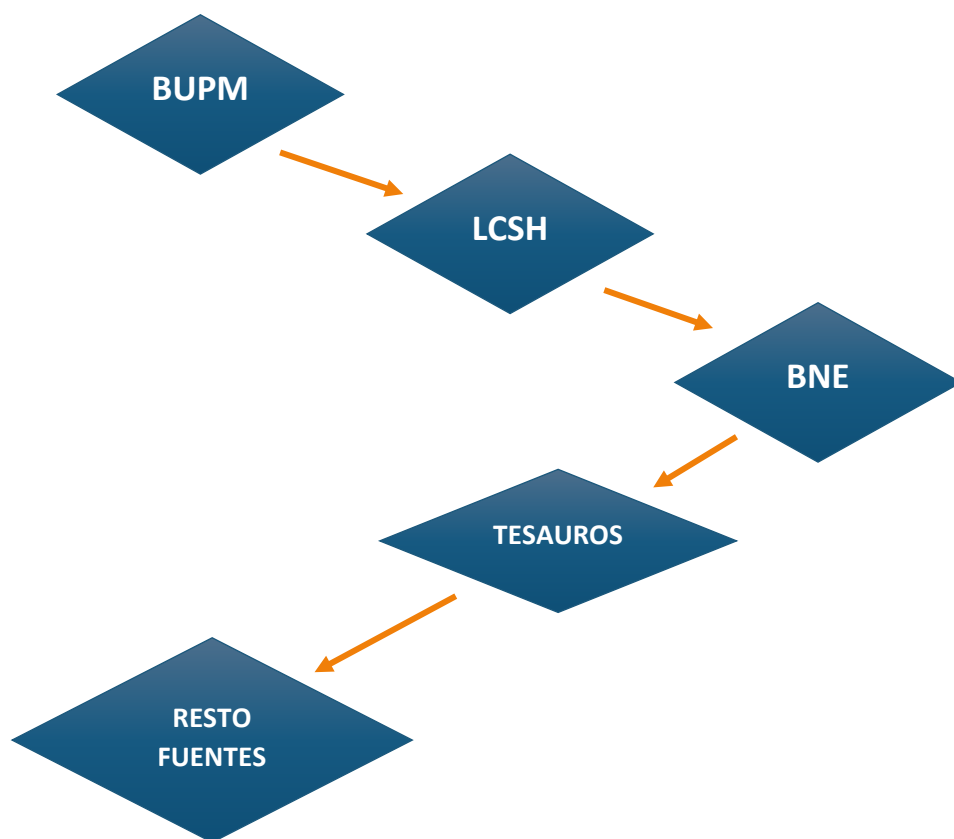


Figura 21 Secuencia de preferencia de fuentes para nuevos conceptos. Fuente: elaboración propia.

5.2.4.1 Representación del modelo de datos BUPM en diagrama UML.

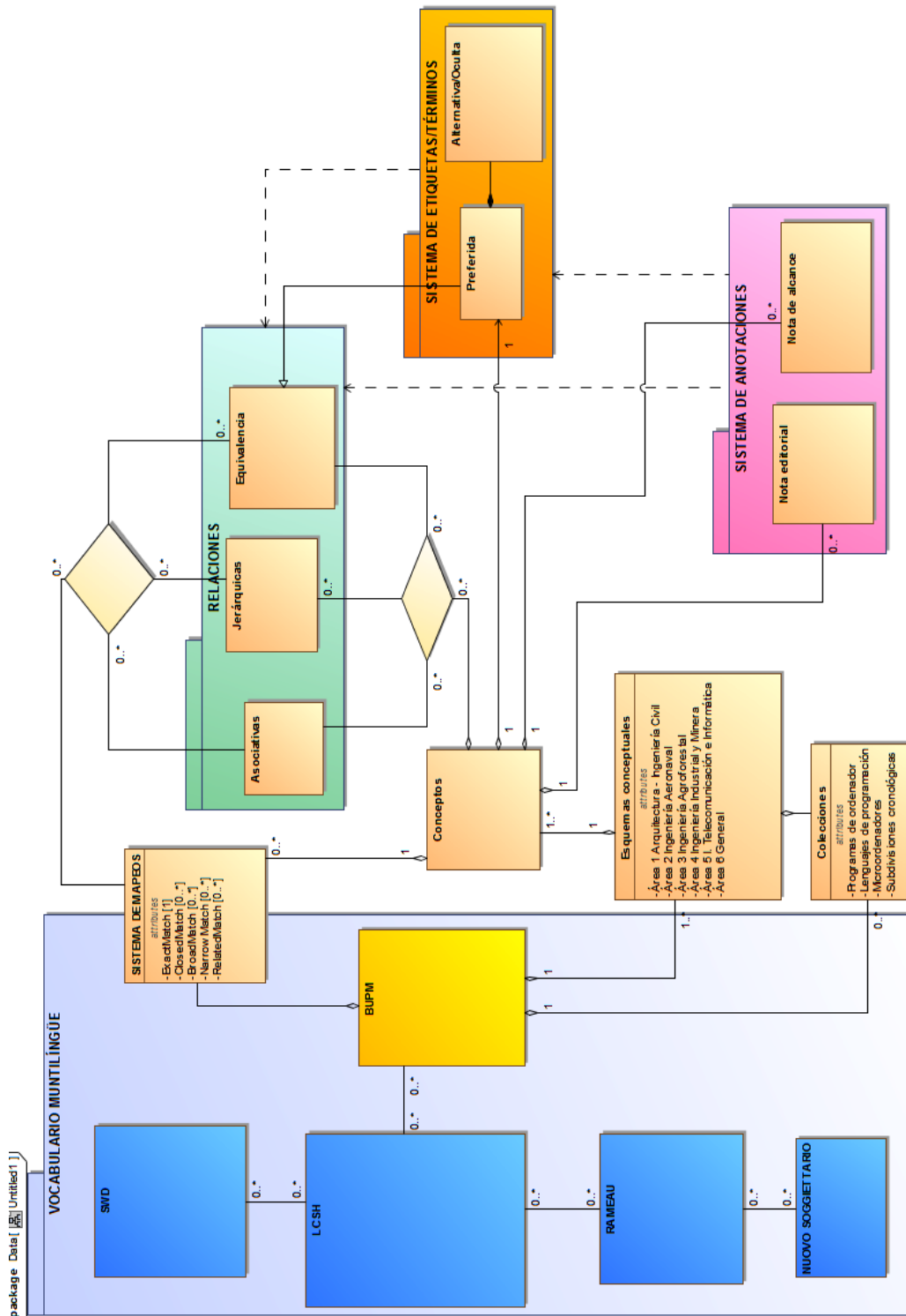


Figura 22 Modelo de datos del vocabulario de materias BUPM. Fuente: elaboración propia.

5.2.5 SELECCIÓN DE LOS CONCEPTOS DEL PROTOTIPO

Los criterios para la elección de los conceptos del prototipo persiguen conseguir algunas mejoras sobre el vocabulario base.

1. Se pretende aumentar y mejorar las relaciones subyacentes en el vocabulario, haciendo más presentes las relaciones asociativas y generando relaciones jerárquicas (no contenidas originalmente) que ofrezcan una mejor expresión semántica. Para ello se escogerán las agrupaciones de conceptos más significativas según los criterios de alcance definidos y que tengan un mayor potencial para definir nuevas relaciones (No se puede constatar el resultado con ningún sistema de recuperación o indización, que verificarían los parámetros de calidad obtenidos).
2. Partiendo de la base de que la migración del vocabulario de materias debe respetar esencialmente el contenido establecido, el ajuste del vocabulario a las normas ISO 25964, o a las recomendaciones de mayor rango, como el Manual de Autoridades BNE, debe ser documentado, justificando la necesidad de ofrecer una nueva versión y haciendo mención en el registro mediante notas de la fuente de procedencia.
3. La modificación del vocabulario no supone cambio alguno respecto a las condiciones de propiedad intelectual ni del vocabulario base, ni de la versión LOD del mismo.
4. Si por necesidades de coherencia semántica, clarificación de relaciones, o ajuste del mapeo con los vocabularios externos de materias hubiera que añadir nuevos encabezamientos al vocabulario, estos tendrán la clasificación de provisionales (susceptibles de someterse a las revisiones antes referidas) y se justificará en las anotaciones del registro tanto su elección como la fuente autorizada utilizada.
5. Para la selección de nuevos conceptos o su forma léxica preferente, se consultarán los vocabularios más reconocidos para cada materia en general.
6. El proceso de selección se dividirá en varias partes:
 - a. La principal, que pretende suministrar conceptos estructurales que permitan construir una versión previa del tesoro.
 - b. La segunda y sucesivas cuyo objetivo será completar y ajustar las estructuras primarias y posteriormente enriquecer el sistema establecido.

(Biblioteca Universidad Politécnica de Madrid, 2010; Harpring, 2012; International Organization for Standardization (ISO), 2011; Leroi et al., 2011; Sánchez-Cuadrado, Morato-Lara, Moreiro-Gonzalez, & Marrero-Linares, 2007)

5.2.6 CONTROL DEL VOCABULARIO

El Manual de autoridades de la Biblioteca Nacional de España indica que los términos (y por extensión los conceptos) deben ser: “.....*únicos, consistentes, normalizados y no arbitrarios, controlados y no libres*”. Cualquier vocabulario adaptado a la web debe seguir estas mismas directrices como fundamento básico de la recuperación y de la indización.

El proyecto que aquí se lleva a cabo gestiona una estructura ya conformada y con unas reglas establecidas de control de vocabulario. Habitualmente los encabezamientos de materia, y en concreto los de la Biblioteca UPM, siguen normas no especialmente diseñadas para la gestión del vocabulario en un entorno digital, lo que sugiere la necesidad de un cierto alineamiento con la ISO 25964 -1 a estos efectos. (Todos los ejemplos formateados en negrita se corresponden a materias integrantes del tesoro de materias).

5.2.6.1 *Control conceptual*

Desde un punto de vista conceptual, la posible la inclusión de nombres propios como materia están bien utilizada, y se corresponde con materias que describen procesos matemáticos u objetos denominados por el nombre de su creador, en ningún caso se han permitido nombres propios sin asociación a la persona, documento u objeto del recurso indizado:

Álgebra de Boole
Análisis de Fourier
Series de Poincaré

Carmen Sandiego

La herramienta de anotación del control conceptual es la nota de alcance, que identifica los límites y especificaciones del concepto necesarias para definir la forma preferente de su expresión terminológica. La naturaleza conceptual de los tesauros en contextos digitales impone la necesidad de establecer relaciones incluso entre notas de alcance. Así es el caso de las denominadas notas recíprocas en las que las referencias a un concepto determinado deben suponer referencias recíprocas en el concepto mencionado. Este mecanismo de vinculación semántica no está soportado por los sistemas de modelado actuales, pero si recogido en la norma

ISO 25964. Cuando se utilice esta relación en el tesauro aquí propuesto se indicará un código que se incluirá idénticamente en ambas notas.

Este tipo de relación se ha establecido entre notas de los conceptos Fruticultura y Productos agrícolas.

5.2.6.2 Control terminológico

El control de homonimias se realiza en el vocabulario mediante adiciones o calificadores que desambiguan los diferentes significados de la forma. La ISO 25964-1 desaconseja su uso en general por su mala integración con aplicaciones, prefiriendo el uso de multitérminos. En cambio sí se recomiendan las adiciones monotérmino con suficiente valor de especificación.

Arqueo (Buques) Visión artificial (Robótica)

No parece adecuado utilizar adiciones con un alto nivel de generalidad. Si se dan las condiciones podrían definirse dos materias con cada uno de los calificadores:

Aviones (Diseño y construcción)

Las adiciones demasiado largas deberían ser sustituidas por monotérminos significativos o incluso eliminar la adición.

Pentaho (Paquete integrado de software) *Pentaho suite (Propuesta)*

Las actualizaciones de términos deberían llevar referencias cronológicas de su creación, cambio de estatus (prefrente-no preferente), etc. Se recomienda la utilización de notas cronológicas (*skos:historyNote*). (*El software de gestión utilizado aquí, no contiene dicha posibilidad, la inclusión de notas históricas deberá hacerse en notas de alcance*).

5.2.6.3 *Formas gramaticales de los términos*

La norma refiere la preferencia de sustantivos o frases sustantivadas. No recomienda, en cambio, los tipos adjetivados y preposicionales (cuestión mucho más fácil de cumplir en inglés que en español). La inclusión de adjetivos en las materias puede generar recuperaciones imprecisas. La norma hace referencia en varias ocasiones a las dificultades de alinear reglas gramáticas en tesauros multilingües. Esta tipología de materias es de uso común en el vocabulario BUPM y de difícil simplificación. En el contexto del Manual de autoridades BNE se indica la función especificativa del adjetivo: para concretar un sustantivo, para las aplicaciones de una técnica, para dar sentido a sustantivos demasiado inespecíficos o para desambiguación.

Tecnología espacial (Adjetival)

Tubos de rayos catódicos (Preposicional)

Software libre

Grabación magnética

Medidas electrónicas

No se recomienda la utilización de adverbios. En el vocabulario no hay materias con componentes adverbiales. Si se admiten en cambio formas nominales del verbo, que en castellano es el participio, aunque en el vocabulario no son comunes.

Diagnóstico asistido por ordenador

El artículo no debe ser utilizado según la norma ISO 25964 salvo por necesidades de recuperación, pues en ocasiones forma parte de una expresión íntegramente considerada concepto. En el vocabulario no se utilizan artículos.

Respecto al uso de mayúsculas el vocabulario sigue las directrices de la norma, utilizando preferentemente minúsculas con mayúsculas en la primera letra. El uso de mayúsculas en todo el término se reduce a acrónimos, siglas o formas que originariamente están íntegramente en mayúsculas. El uso de caracteres no alfabéticos también está desaconsejado para evitar conflictos en las búsquedas.

Las especiales características de este vocabulario, con una mayoría de conceptos de carácter científico-técnico, obligan a la inclusión de caracteres no recomendados. Cabe también la posibilidad de incluir como términos no preferentes las mismas expresiones sin símbolos.

PAN (Programa de ordenador)
OS/2 (Sistema operativo)
PL-1 (Lenguaje de programación)
QSB+ (Programa de ordenador)

Según el Manual de autoridades de la BNE, el uso de plurales o singulares en los términos (conceptos) depende del uso lingüístico. Cabe la posibilidad de términos que expresen cantidades, de conceptos abstractos o concretos e incluso una forma singular o plural de una misma palabra puede tener diferentes significados. En general se indica el singular para conceptos abstractos, propiedades, procesos, creencias, etc., y se recomienda el plural en cambio para grupos de seres vivos o personas, partes del cuerpo, etc. Estas normas están recogidas íntegramente en las pautas para encabezamientos de materia de la BUPM. La norma ISO 25964 hace referencia a conceptos contables y no contables, teniendo en cuenta las diferencias idiomáticas. En general los contables se introducen en plural y los no contables en singular.

Refino
Socialismo
Taoísmo
Psiquiatría
Refracción

Riñones
Árboles
Puentes
Aves acuáticas

5.2.6.4 Criterios para la elección de términos preferentes

Las formas preferentes en listas de encabezamientos de materia se refieren a los propios encabezamientos autorizados. Las formas no autorizadas del término se asimilan a las no preferentes. Esta circunstancia es importante a la hora de migrar la estructura típica de LEM, pues se trata, entre otras cosas, de trasladar la estructura de equivalencia de un vocabulario a otro. En principio y como norma general, la BUPM sigue la directriz de la BNE para escoger el encabezamiento de materia, el “Principio de autoridad literaria o bibliográfica” que supone que se escogerá el término que se suele usar habitualmente en la disciplina, en combinación con el “Principio de autoridad del usuario” cuyas preferencias deben constar a la hora de la elección. En

el caso que aquí nos ocupa, los conceptos nuevos y añadidos o con cambio de estatus que incluyan términos no preferentes serán justificados en anotación, reflejando la fuente de referencia del área.

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/1284> a skos:Concept ;
  skos:prefLabel "Comunicación móvil"@es ;
  skos:closeMatch <http://id.loc.gov/authorities/sh85086371#concept> ;
  skos:notation "621.395"@es ;
  skos:altLabel "Sistemas de comunicaciones móviles"@es ;
  skos:narrower
  <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/1288> .
```

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/1288> a skos:Concept ;
  skos:prefLabel "Comunicación móvil 3G"@es ;
  skos:scopeNote "Nota editorial. Cambio de estatus del concepto.
Comunicación móvil 3G cambia de no autorizado a autorizado y se asigna
a la jerarquía: Sistemas de telecomunicación/Comunicación móvil. Se
desvincula la referencia cruzada con el concepto Sistemas universales
de comunicación móvil."@es ;
  skos:notation "621.395"@es ;
  skos:altLabel "Comunicación de tercera generación"@es ;
  skos:broadMatch <http://id.loc.gov/authorities/sh85086371#concept> ;
  skos:relatedMatch <http://id.loc.gov/authorities/sh96011791#concept> ;
  skos:broader
  <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/1284> .
```

La norma ISO 25964 hace referencia en primer lugar a las cuestiones ortográficas, haciendo referencia a la selección de la forma común. Las formas con incorrecciones ortográficas están prohibidas como términos preferentes y deben identificarse como formas mal escritas (son candidatas obvias para añadirse mediante una *hiddenLabel*. El vocabulario de materias BUPM no contiene términos bajo formas ortográficamente incorrectas.

Los términos prestados de otras lenguas deben ser designados preferentes según las normas anteriores, en cualquier caso pueden ser añadidos como no preferentes.

Art déco

Altavoces

Baffles

Bádminton

Volante (Juego)

El vocabulario contiene algunos ejemplos de conceptos prestados que no son preferentes pero se incluyen dado el volumen de documentos indizados con ellos.

Arquitectura orientada a servicios (Informática)

Service-oriented architecture (Informática)

Las transliteraciones deben ser efectuadas según reglas establecidas que deben constar expuestas en las referencias del tesoro.

Chikung “Trasliteración del lenguaje chino del concepto Tai Chi.”

Los neologismos y términos de jerga deben ser incluidos si ayudan a la recuperación. En un vocabulario de estas características con una cantidad apreciable de conceptos que designan objetos o fenómenos de las nuevas tecnologías, es preciso mostrar agilidad y criterio a la hora de incorporar neologismos.

CD-ROMs

DVD

Benchmarking

Weblogs

Skinheads

Arquitectura High-Tech

Otras formas de preferencia son las de los nombres comunes sobre las marcas comerciales que también designan el concepto. Respecto a las formas populares de los términos o sus nombres científicos, debe asignarse una política acorde al tipo y alcance del tesoro. El vocabulario aquí tratado contiene una sustancial cantidad de materias correspondientes a procesos y objetos técnicos diversos relacionados con la ingeniería, escogidos habitualmente por la forma más común presente en la literatura.

Pteridophyta

Angiospermas

Respecto a las abreviaturas y los acrónimos, la norma refiere la preferencia general sobre la forma completa, pero admitiendo las formas abreviadas si son las más comúnmente conocidas y utilizadas con predominio sobre la forma completa. En el vocabulario se encuentran ambos tipos de formas.

Redes de área extensa (Redes de ordenadores)

WAN (Redes de ordenadores)

Redes digitales de servicios integrados

RDSI

SIDA

Síndrome de inmunodeficiencia adquirida

Sistemas MIMO

Sistemas Multiple Input Multiple Output

UML (Informática)

Unified Modeling Language (Informática)

Respecto a los nombres propios de personas y organizaciones, la norma remite a la utilización de otros vocabularios de autoridades específicos para designar formas normalizadas del nombre, o las normas de catalogación del centro. En el caso del vocabulario aquí tratado, y dado que se trata de uno más entre los listados de autoridades de la biblioteca, se debería impedir su inclusión indiscriminada, excepto para los casos ya apuntados arriba: procesos, objetos, técnica, etc., cuya denominación más conocida incluye el nombre propio del autor, intérprete, etc.

Álgebras de Banach

Cemento Portland

Sistemas de Steiner

Respecto a los lugares se pueden utilizar indistintamente teniendo en cuenta el más utilizado por la comunidad que utiliza el tesoro. No son habituales los nombres propios geográficos y en general se introducen en conceptos complejos bajo la forma oficial. Estos pueden integrar también los propios encabezamientos geográficos.

Denominación de origen Valencia

Denominación de origen Campo de Borja

5.2.6.5 Criterios para la formación de encabezamientos complejos

La migración desde una lista de encabezamientos de materia a un tesoro no sería una tarea difícil si no fuera por la distinta filosofía de los conceptos complejos en las LEM y en los tesauros. Las listas de encabezamientos de materia persiguen especificar los temas de un documento, cuanto más precisamente lo hagan, mejor recuperabilidad del recurso en cuestión. A veces expresar una materia con una sola palabra no es posible o no es eficaz, por lo que se deben utilizar expresiones más complejas formadas por uniones de conceptos mediante conjunciones o preposiciones, o formando encabezamientos complejos mediante la adición de subdivisiones, o mediante el mecanismo denominado precoordinación. Las normas de tesauros, en cambio, consideran necesario simplificar los conceptos, intentando minimizar su complejidad dado que para la recuperación cabe siempre la posibilidad de unirlos coyunturalmente, es la denominada postcoordinación.

Estos diferentes enfoques debilitan el modo de expresión de un vocabulario de materias en Linked Data, pues normativamente la precoordinación apenas existe y los mecanismos que se utilizan en algunos casos (MADS) no recogen completamente su operatividad. Efectivamente puede argumentarse que el contexto digital no necesita de mecanismos como los que tienen las LEM para organizar el conocimiento, pero no se trata de dilucidar aquí si es mejor un vocabulario que otro, sino de poder expresar en toda su amplitud las capacidades y peculiaridades de cada uno, conformando un abanico de posibilidades de organizar la información. En otras partes de este trabajo se han presentado modelos para llevar a cabo la precoordinación (SKOS-XL y MADS), éstos sistemas no tienen posibilidad de aplicación automática en el software de gestión, por lo que serán aplicados de modo manual y selectivo a modo de ejemplo a algunos grupos de conceptos.

Para la BNE existen tres grupos fundamentales de encabezamientos complejos: los adjetivales, ya tratados más arriba, los formados por conceptos unidos mediante conjunciones, preposiciones y adverbios, y los encabezamientos complejos con la adición de subdivisiones. Los compuestos por conjunciones, preposiciones y adverbios son comunes en la lista de encabezamientos BUPM, su migración al tesoro no requerirá la división de conceptos complejos.

Sistemas de planos acotados
Ciencia y filosofía
Canales de riego
Cálculo relacional difuso

La norma ISO 25964 recomienda la división de los conceptos complejos en los supuestos en los que esto no perjudique a la indización o la recuperación. El concepto complejo en gran parte de las ocasiones está formado por un concepto principal y otro u otros que los especifican,

reconociendo que esta especificidad mejora la recuperabilidad de los recursos. Curiosamente este argumento es el que fundamenta la precoordinación.

Clima húmedo Fibras de madera

En cualquier caso se puede considerar la pauta general de admisión de conceptos complejos desde el punto de vista de la norma, teniendo en cuenta que es posible que la automatización de los sistemas de indización y recuperación hagan irrelevante la posibilidad o no de componer conceptos.

Son varias las cuestiones que tienen que verificar los protocolos para la efectiva admisión de un concepto complejo:

1. Si tiene un nivel de uso suficiente.
2. Si mantener la complejidad mejora la recuperación de la información.
3. Si las búsquedas son menos efectivas en contextos de densidad de conceptos complejos relacionados.
4. El aumento de la complejidad en conceptos con más de dos componentes.
5. Efectos en los tesauros multilingües de los conceptos complejos (por añadidura los mapeos son más complicados).

No en todos los casos la complejidad conceptual es un inconveniente, curiosamente la misma norma hace referencia al caso de conceptos complejos de naturaleza precoordinada que son eficaces en la recuperación, se trata del caso de utilización del tesoro por especialistas en la materia, que saben aprovechar la riqueza expresiva de las LEM. En cambio desaconseja su utilización para tesauros de índole general.

En el caso de aceptación del concepto complejo debe representarse bajo la forma de término preferido, y también los términos que lo componen han de ser almacenados como términos preferentes, uno de ellos como genérico de la forma compuesta, el otro como relacionado con el genérico. En definitiva la referencia del tesoro debe contener una política definida para regular la admisión o no de conceptos complejos, reglas para establecerlo, para dividirlo etc., todo ello para establecer un marco coherente de gestión del vocabulario.

La problemática de los conceptos complejos se ha manifestado en el contexto del vocabulario BUPM a la hora de efectuar mapeos. Los diferentes supuestos encontrados se han gestionado de acuerdo a la segunda parte de la norma ISO 25964 que contempla procesos para la conversión de conceptos de diversa complejidad, tanto en el vocabulario fuente, como en el objetivo, allí se han expuesto los ejemplos pertinentes.

5.2.6.6 Uso de las adiciones

El vocabulario de materias de la BUPM hace un uso intensivo de las adiciones y lo hace en concordancia con las pautas de la BNE, es decir, para aclarar conceptos por su complejidad o por su excesiva generalidad. El término utilizado como adición debe estar normalizado tendiendo hacia términos genéricos:

TrueSpace (Programa de aplicación)
Plantas (Botánica)
Asignación de recursos (Informática)
AMS-LaTeX (Programa de ordenador)

También es útil utilizar adiciones cuando necesitamos desambiguar homónimos:

Árboles
Árboles (Teoría de grafos)

Arena
Arena (Programa de ordenador)

La especificación de expresiones geográficas no se contempla en el vocabulario BUPM (tampoco en los encabezamientos geográficos que no son objeto de este trabajo). Algunas referencias a zonas geográficas, como las denominaciones de origen no se incluyen como adición dado que la especificación geográfica es la palabra principal del encabezamiento.

Denominación de origen Rioja
Denominación de origen Rueda
Denominación de origen Utiel-Requena

(Biblioteca Nacional de España, 2014c; Biblioteca Universidad Politécnica de Madrid, 2010; International Organization for Standardization (ISO), 2011).

5.2.7 GESTIÓN DE LAS RELACIONES

5.2.7.1 *Relaciones entre términos preferidos y no preferidos*

El Manual de autoridades BNE califica a las relaciones de equivalencia como aquellas que se establecen entre un encabezamiento autorizado, y aquellos encabezamientos alternativos que designan el mismo concepto y que deben ser referenciados al encabezamiento principal. Las pautas sobre materias de la BUPM siguen fielmente este planteamiento estableciendo un encabezamiento principal (al que denominan descriptor) y opcionalmente varios alternativos vinculados por la fórmula “Usado por”. (*Todos los ejemplos de jerarquías se corresponden con la estructura real establecida en tesauro*).

Ingeniería de sistemas

UP *Planificación de sistemas*

En el ámbito de los tesauros, la norma ISO 25964 define la equivalencia en términos de etiquetas preferentes (que asimilaremos al concepto de encabezamiento principal) y no preferentes (que harán referencia a las etiquetas alternativas). En este caso, la utilización de las referencias “Use” y “Usado por” quedan circunscritas a las posibles versiones impresas pues en el contexto digital, dichas referencias se infieren de la información del registro de cada concepto donde SKOS asigna una *prefLabel* y opcionalmente las *altLabel* e *hiddenLabel*. Las principales formas de relaciones de equivalencia se establecen por: sinonimia, cuasinonimia, equivalencia de los términos genéricos y equivalencia compuesta. Podemos hacer un recorrido observando la adecuación del vocabulario a las recomendaciones de la norma sobre equivalencia:

5.2.7.1.1 Sinonimia

Los casos donde se plantea la sinonimia pueden ser varios, por ejemplo por diferente origen idiomático:

Encaminadores (Redes de ordenadores)

Routers

Hardware

Soporte físico de ordenadores

Palabras populares frente a su forma científica:

ADN

Ácido timonucleico

Nombres comunes y marcas comerciales:

Office

Microsoft Office

Conceptos emergentes:

Teléfonos inteligentes

Smartphone

Conceptos actuales frente a conceptos en desuso:

Middleware

Soporte intermedio de ordenadores

Formas abreviadas y acrónimos:

ODBC

Open Database Connectivity

MP3

MPEG Audio Layer 3

5.2.7.1.2 Quasi-sinonimia

Puede referirse a términos con un grado de complementariedad muy alto o por otra parte los antónimos. En cualquier caso para un mejor alineamiento a las recomendaciones de la norma se debe tener en cuenta que el alcance del vocabulario puede permitir quasi sinónimos con campos semánticos afines, pero también es posible que si el dominio del tesoro es muy específico, los

candidatos a quasi sinónimos deben conceptuarse por separado. En el vocabulario BUPM de materias no se utiliza la antonimia.

Transporte colectivo

Transporte de pasajeros

Cubiertas laminares

Estructuras laminares

Edificios inteligentes

Hogar digital

Las relaciones entre términos generales y específicos (fuera del ámbito jerárquico) como equivalentes no tienen un tipo representativo en el vocabulario de materias, tampoco se muestran equivalencias compuestas.

5.2.7.2 Relaciones entre conceptos. Relaciones jerárquicas y asociativas.

Aunque conceptualmente es posible expresar relaciones jerárquicas en el contexto de las listas de encabezamientos de materia, en no pocas ocasiones estas relaciones no aparecen o aparecen pseudo representadas a través de las relaciones asociativas. Es el caso del vocabulario con el que trabajamos aquí, donde no existen explícitas relaciones jerárquicas, pero sí implícitas. El Manual de autoridades de la BNE describe las relaciones jerárquicas como una relación de subordinación o superordinación entre conceptos, al igual que la ISO 25964 que define la clase *HierarchicalRelationship* como expresión de las mismas. Todos los tipos de relaciones subsiguientes se modelan de acuerdo a la norma antes citada. Se contemplan varios tipos:

1. Relación genérica: un concepto es clase con respecto a otro, el cual es miembro a su vez de esa clase. Es la tipo de relación jerárquica más habitual y de conversión más sencilla. La especificación de la ISO 25964 a estos efectos recoge un etiquetado específico de tipo para las jerarquías genéricas: BTG, *broader term generic* y NTG, *narrower term generic*) :

BTG Bases de datos

NTG Bases de datos deductivas

NTG Bases de datos distribuidas

NTG Bases de datos estadísticas

2. Relaciones parte-todo: donde el concepto forma parte del alcance del concepto general. Nos ceñimos al marco establecido por la ISO 25964 que posibilita únicamente este tipo de

relaciones en cuatro casos principales: órganos del cuerpo, localizaciones geográficas, disciplinas o áreas del conocimiento y estructuras jerárquicas sociales (etiquetado ISO BTP *broader term partitive* y NTP *narrower term partitive*). El vocabulario BUPM hace un uso intensivo de este tipo de relaciones sobre todo en las divisiones de las ciencias.

BTP Ciencia

NTP Matemáticas
NTP Biología
NTP Física
NTP Geología

3. Relaciones enumerativas: que encabezan con una clase superior seguida de ejemplos o instancias de la misma, habitualmente expresados como nombres propios. Se introduce un ejemplo con los encabezamientos geográficos, aunque no están contenidos en este trabajo (etiquetado BTI *broader term instantial* y NTI *narrower term instantial*).

BTI Provincias -- España

NTI Barcelona
NTI Sevilla
NTI Badajoz
NTI Valencia

4. Relaciones polijerárquicas son aquellas en las que un concepto puede estar bajo la estructura jerárquica de varios conceptos a la vez. En el caso particular que nos ocupa y debido a la estructuración del tesoro en varios *ConceptScheme* o *Thesaurus* según terminología de la ISO 25964, se establecen relaciones polijerárquicas en algunos casos. De facto, el listado de materias BUPM se encuentra temáticamente dividido por áreas de conocimiento que cubren los diferentes ámbitos de las ingenierías que se imparten en la UPM. Esa división se ha respetado en el tesoro (aunque en algunos casos se ha decidido “importar” conceptos de un área a otra, justificándolo en *EditorialNote*). Este tipo de estructuración ha provocado que las relaciones polijerárquicas aparezcan en cierto número.

BT Industria

NT Sociología industrial

BT Sociología

NT Sociología industrial

BT Ingeniería

NT Ingeniería electrónica

Thesaurus (Telecomunicaciones e Informática)



Ingeniería electrónica isTopConceptOf Telecomunicaciones e Informática

En el primer caso estamos ante una relación polijerárquica establecida con sendos conceptos en diferentes *Thesaurus*. En el segundo, el concepto “Ingeniería electrónica” depende del *TopConcept* “Ingeniería” y es a su vez *TopConcept* del *Thesaurus* “Telecomunicaciones e Informática”. En el modelado SKOS se declaran que los conceptos polijerárquicos en ambos “*ConcepScheme*”, mediante el establecimiento manual de la propiedad `skos:broader`, para cada uno de sus conceptos genéricos.

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/487> a skos:Concept ;
  skos:broader <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/472>
  skos:broader <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/486>
  skos:prefLabel "Sociología industrial"@es ;
  skos:scopeNote "Nota editorial: concepto provisional"@es , "Nota de
  alcance: polijerarquía"@es ;
  skos:notation "316.334.22"@es .
```

La pauta de referencia en la ordenación jerárquica ha sido la establecida por el propio alcance de los conceptos contrastado con las listas de encabezamientos de materia de la Biblioteca Nacional de España y de la Library of Congress Subject Headings. En algunos casos se han utilizado términos provisionales del listado de materias BUPM para completar ramas jerárquicas, en cuyo caso se ha especificado tal circunstancia en nota editorial. La propia norma faculta al editor del tesoro a construir subgéneros de relaciones si lo considera necesario para responder a las necesidades de indización y de búsqueda de la comunidad usuaria del tesoro. Establecer una estructura tan específica para un vocabulario de estas características podría mejorarlo en algunos aspectos, incluso la utilización de facetas en grandes áreas como la Informática, donde una división como la siguiente podría tener sentido. Pero el incremento de la dificultad y complejidad de las jerarquías y la imposibilidad de expresarlo mediante el gestor de tesauros, desaconsejan su utilización en este caso:

Informática

<Agentes>

Informáticos

.....

<Aplicaciones>

Análisis de datos

.....

<Software>

Lenguajes de programación

.....

<Hardware>

Arquitectura de ordenadores

.....

<Sistemas informáticos>

Sistemas expertos

.....

5.2.7.3 Relaciones asociativas

Según la ISO 25964 la relación asociativa supone una vinculación semántica entre los conceptos que posibilita nuevas opciones a efectos de la recuperación o la indexación, recalcando la gran importancia que tiene disponer estas relaciones en los tesauros. Este tipo de relación supone que la utilización de un concepto implica a sus conceptos relacionados en los espacios semánticos utilizados por los usuarios de los vocabularios y esto sucede porque la relación asociativa es siempre recíproca. Ya se ha comentado en este trabajo que en el vocabulario de origen se utilizan relaciones asociativas que en realidad pueden ser jerárquicas, por lo que se ha procedido a convertir en algunos casos las relaciones preestablecidas. La tipología de relaciones asociativas subyacentes se ha alineado con las pautas de la BNE y de la ISO 25964:

Significados solapados:

BT Automatización

- TE Automatización de bibliotecas
- TE Automatización de oficinas
- TR Aspectos sociales de la automatización
- TR Teoría de autómatas

Una disciplina y sus objetos de estudio:

Informática

- TR Ordenadores
- TR Sistemas informáticos

Proceso y su instrumento:

Computabilidad

- TR Ordenadores

Una acción y su producto

Lenguajes de programación

- TR Programas de ordenador

Refino

- TR Gasolina

Una acción y su contenedor

Minería de datos

- TR Bases de Datos

Según la ISO 25964, algunos tipos de relación pueden establecerse de modo asociativo o de modo jerárquico dependiendo de las características del vocabulario. En el ejemplo siguiente del

vocabulario se escoge la relación asociativa por la importancia independiente de ambos conceptos en un esquema que pretende agrupar materias de la Ingeniería Naval, donde el concepto “buque” tiene demasiado peso semántico:

Construcción naval
TR Buques

Una sustancia derivada de otra

Hierro
TR Acero

(Biblioteca Nacional de España, 2014c; Biblioteca Universidad Politécnica de Madrid, 2010; International Organization for Standardization (ISO), 2011)

5.3 PREPARACIÓN DEL VOCABULARIO PARA LA INTEROPERABILIDAD

5.3.1 IDENTIFICACIÓN DE DATOS. INTERNATIONALIZED RESOURCE IDENTIFIERS

Aunque el gestor de tesauros no nos permite crear libremente los IRI que identifican los datos, parece conveniente, para la adaptación normativa a la Norma Técnica de Interoperabilidad, ofrecer la propuesta de modelado del sistema de identificación de recursos.

La NTI establece como obligatorio, la utilización de estos identificadores para la representación de recursos sin ambigüedades, de modo persistente y con garantías de procedencia (Ministerio de Hacienda y Administraciones Públicas, 2013). Para facilitar el trabajo a los agentes reutilizadores de los datos de las AAPP, se requiere que estos identificadores se modelen mediante un sistema común que siga estas pautas:

1. Utilización de protocolos HTTP o HTTPS.
2. Ofrecer al reutilizador la posibilidad de recibir el recurso (o dato) en el formato adecuado según la petición (La negociación de contenido no ha podido ser comprobada en la plataforma PoolParty, se sobreentiende que el sistema podrá atender tanto peticiones manuales como de agentes automáticos). La identificación debes estar establecida de tres modos diferentes: para humanos (HTML), para procesos automáticos (RDF) y la IRI abstracta que identifica el recurso.
3. Se establecerá un modelo de identificador según las pautas del Anexo II NTI.

4. El identificador no sólo debe proporcionar información suficiente del recurso o dato, sino también de su procedencia.
5. El identificador no debe contener información técnica sobre la implementación del recurso (Ministerio de Hacienda y Administraciones Públicas, 2013), se evita así que cualquier migración técnica suponga el desfase de todo el sistema de identificadores.

La definición del esquema IRI para la Norma Técnica de Interoperabilidad tiene estos componentes prioritarios:

Base: que ofrece la información de procedencia y tiene carácter obligatorio, en nuestro caso.

`http://www.upm.es/biblioteca`

Carácter de la información: expresa la naturaleza del recurso, está normalizado y es obligatorio, al ser un sistema de organización del conocimiento escogemos *kos*.

`http://www.upm.es/biblioteca/kos`

El elemento Sector, requiere la utilización de un identificador de actividad representado en una taxonomía. Su naturaleza es opcional y no es utilizado aquí.

El elemento dominio permite un acercamiento descriptivo mayor a la naturaleza del recurso. En este caso empleamos la referencia “sh” para identificar que estamos ante un vocabulario de materias. Es un elemento opcional.

`http://www.upm.es/biblioteca/kos/sh`

Los conceptos específicos son expresados mediante el código asignado actualmente a cada materia (este código es el asignado por el sistema de bases de datos actual de la lista de materias) y se utilizan aquí como identificadores abstractos:

`http://www.upm.es/biblioteca/kos/sh/XX125453`

El modelado de los IRI para solicitudes automáticas y de interfaz humana son respectivamente:

`http://www.upm.es/biblioteca/kos/sh/XX125453.ttl`

<http://www.upm.es/biblioteca/kos/sh/Internet.html>

La NTI no olvida la fragilidad de los sistemas de identificación en Internet y regula una serie de pautas para mantener la coherencia y favorecer un más fácil mantenimiento en el futuro de los IRIs:

1. Utilizar identificadores alfanuméricos representativos y semánticos.
2. Utilizar minúsculas excepto para las clases.
3. Evitar signos de puntuación, acentos etc.
4. Eliminar artículos y conjunciones en su caso.
5. Utilizar guiones para unir palabras.

(Ministerio de Hacienda y Administraciones Públicas, 2013)

5.3.2 ESTABLECIMIENTO DE RELACIONES SEMÁNTICAS CON OTROS VOCABULARIOS

Difícilmente se puede hablar de Linked Data si no establecemos vínculos con otros conjuntos de datos. En el caso que nos ocupa, se ha pretendido mejorar el contenido del tesoro de materias BUPM estableciendo mapeos con los principales vocabularios de materias compartidos en la Web de datos. Este proceso, aunque beneficioso, es de una gran complejidad, y no por los parámetros técnicos, como cabría esperarse, sino por los ajustes semánticos, difíciles de establecer en el propio idioma, cuanto más entre varios vocabularios de materias en diferentes lenguajes. El objetivo primordial, conseguir un producto multilingüe, en el que el inglés no predomine y se pueda aprovechar la riqueza semántica de los diferentes idiomas.

La planificación sobre qué vocabularios escoger para los mapeos no ha requerido un gran esfuerzo, se trata de enlazar con vocabularios de valor en el sector de los encabezamientos de materia, y a nuestro juicio estos son principalmente la LCSH, RAMEAU, SWD y Nuovo Soggettario. Todos ellos tienen un dominio enciclopédico que permite abarcar con facilidad los campos semánticos de las materias BUPM. Otro punto relevante ha sido la comprobación de las licencias de uso de los diferentes vocabularios:

Library of Congress Subject Headings: Public Domain by Attribution. (Los requisitos legales para usar el vocabulario obligan a hacer explícito el origen del vocabulario, es decir no se puede establecer una interfaz en la que se confunda el origen de los datos que la Library of Congress ofrece).

RAMEAU: La licencia de uso del vocabulario de materias de la BNF requiere la atribución de su procedencia y es libre para utilizaciones no comerciales. En caso contrario está sometida la utilización a tarifa.

SWD: Public Domain Zero. Licencia totalmente abierta sin ninguna restricción.

Nuovo Soggettario: Attribution 2.5 Generic (CC BY 2.5).

Se comprueba que en ningún caso hay impedimentos legales para la utilización de los vocabularios en el contexto de este proyecto. El mero establecimiento de mapeos no encubre la procedencia de los datos vinculados, además de reconocer dicha autoría en el fichero VoID de metadatos.

Otro punto importante que hay que considerar es el propio ajuste semántico. La literatura más acreditada en el tema aconseja que el establecimiento de alineamientos semánticos, sea efectuado por expertos en la materia, lo que aquí se ha suplido con un análisis de los mapeos ya establecidos en los vocabularios antes apuntados en conjunto con la investigación personal. La recomendación también indica la necesidad de incluir un mapeo, al menos, por término preferente, cuestión que aquí se intenta respetar únicamente en los dos primeros niveles, un mayor nivel de profundidad en los mapeos hubiera dilatado el proyecto fuera de unos límites razonables.

Como en otras partes de este proyecto, el gestor del vocabulario establece ciertas limitaciones a la hora de mapear conceptos. PoolParty ofrece una interfaz de consulta semiautomática de las materias de la Library of Congress (se requiere traducción de los conceptos, con los riesgos que ello supone), pero carece de la posibilidad de incrementar esos mapeos a otros vocabularios de materias, ni siquiera efectuando los mapeos de modo manual. Sí establece, en cambio, todas las tipologías de mapeos modeladas por SKOS, lo que permite desarrollar una estructura rica en relaciones si se efectúa el análisis adecuado. También permite establecer mapeos muy interesantes de modo similar, con vocabularios centrales en el mundo de los datos vinculados como son Geonames o Dbpedia.

5.3.2.1 Procedimiento de mapeo del vocabulario de materias.

Para establecer los mapeos entre los vocabularios referidos se van a seguir las pautas reguladas en la norma ISO 25964 -2 cuyo objetivo es analizar los elementos y características que participan en la interoperabilidad de vocabularios y ofrecer recomendaciones para establecer y mantener mapeos entre ellos (International Organization for Standardization (ISO), 2012).

Define la norma el mapeo, como la vinculación entre conceptos de diferentes vocabularios, teniendo en cuenta sus distintas especificaciones y diferencias de alcance. La norma ofrece,

asimismo, diferentes modelos de estructura del sistema de mapeos, cuyo tipo más adecuado para el proyecto es el de vinculado directo, formula recomendada para establecer mapeos entre vocabularios de diferente estructura, ejecutándose en este caso, concepto a concepto. Este modelo también, es el más adecuado para mapeos múltiples, en el que se van a ver involucrados más de un vocabulario. En el desarrollo inferior introducimos a modo de ejemplo el modelado SKOS para el mapeo multilingüe establecido.

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/546> a skos:Concept ;
    skos:prefLabel "Edad de hierro"@es ;

    skos:exactMatch <http://id.loc.gov/authorities/sh85068153#concept> ;
    skos:narrowMatch <http://id.loc.gov/authorities/sh2008124152#concept> ;
    skos:exactMatch
    <http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb11946304b> ;
    skos:closeMatch <http://d-nb.info/gnd/4014102-0> ;
    skos:closeMatch <http://purl.org/bnfc/tid/11176> ;

<http://id.loc.gov/authorities/sh85068153#concept> a skos:Concept ;
    skos:prefLabel "Iron age"@en .

<http://id.loc.gov/authorities/sh2008124152#concept> a skos:Concept ;
    skos:prefLabel "Iron age--Europe"@en .

<http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb11946304b>
a skos:Concept;
    skos:prefLabel "Âge du fer"@en .

<http://d-nb.info/gnd/4014102-0> a skos:Concept;
    skos:prefLabel "Eisenzeit"@de .

<http://purl.org/bnfc/tid/11176> a skos:Concept;
    skos:prefLabel "Civiltà del ferro"@it .
```

El tesauro de materias BUPM ha sido ajustado en lo posible a la norma, aunque han existido dificultades debido a las específicas características del vocabulario y de las herramientas (gestor) empleadas para su construcción.

1. Una vez definida la estructura del vocabulario fuente, con expresión de las principales categorías que describen la semántica del tesauro, se procede a la asignación de vínculos para los niveles jerárquicos 1º y 2º, efectuando dicho proceso en primer lugar entre los encabezamientos de materia de la BUPM y de la LCSH. Para el ajuste semántico se ha extraído si existe, la información de los campos MARC 670 y en el caso de no estar descrito ese campo se decide acudir a fuentes autorizadas que incluyan mapeos con conceptos de la LCSH, fundamentalmente la lista de encabezamientos de materia de la BNE, la cual está alineada con la lista americana, aunque no mediante técnicas Linked Data. En los casos en

los que la información de la BNE no aparece, o no contiene el concepto a mapear, o tiene una expresión diferente, se acude a la propia LCSH para intentar seleccionar el concepto de las informaciones que aparecen en notas o a través (y esta ha sido una técnica muy útil) explorando entre los conceptos genéricos del concepto a mapear.

2. La dirección de mapeos no es recíproca, estableciéndose relación únicamente en la dirección vocabulario UPM a LCSH.
3. La sistemática de mapeos se establece dominio o dominio y dentro de ellos concepto a concepto.
4. El procedimiento de asignación de mapeos ha sido efectuado en primera instancia de modo semiautomático, pues PoolParty permite la aplicación de vínculos con la LCSH mediante una interfaz integrada de búsqueda. Los mapeos con el resto de vocabularios se efectúan manualmente y con las mismas características que el anterior, también para los dos primeros niveles jerárquicos y con la siguiente secuencia: RAMEAU-SWD-NUOVO SOGGIETARIO. Aunque esta información no figura en la interfaz de visualización de PoolParty, se puede acudir para su consulta al fichero con los datos formateados alojado en CKAN.
5. La documentación de los mapeos no puede establecerse en PoolParty por carecer de la propiedad *skos:editorialNote*. Para solventar este inconveniente, todos los supuestos de documentación, incluidos los mapeos, se incluyen en nota de alcance haciendo referencia en su contenido de que se trata de una nota de edición.

Una de las estructuras más interesantes valoradas es la denominada estructura en *hub*, para ello se requiere que uno de los vocabularios se establezca como puente o conmutador central, paso obligado de los mapeos entre todos los demás vocabularios.

Se precisa que uno de los vocabularios tenga un dominio lo suficientemente amplio como para abarcar todo el posible alcance de los demás. En el caso que nos ocupa dicha estructura podría ser definida del siguiente modo:

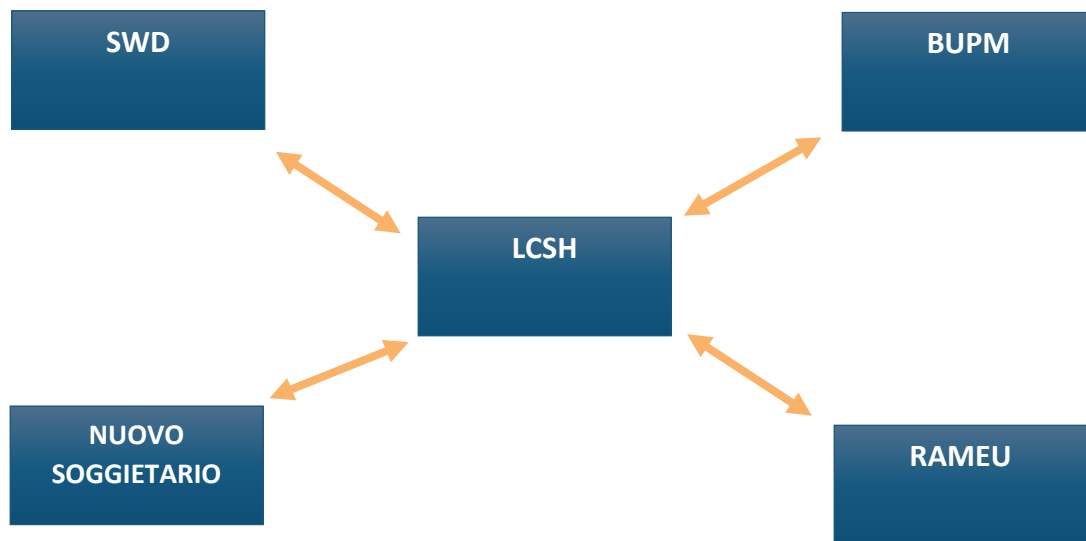


Figura 23 Estructura de mapeo en hub. Fuente: elaboración propia.

Durante la indización o la búsqueda, los conceptos requeridos en cualquiera de los vocabularios satélites se vinculan (convierten como dice la norma) con el vocabulario *hub* y a través de éste con cualquiera de las demás. Estas posibilidades de mapeo se establecen a nivel de aplicación, y su expresión a través del modelado no es posible (International Organization for Standardization (ISO), 2012).

5.3.2.2 Establecimiento de mapeos de equivalencia, jerárquicos y asociados

Las relaciones que se establecen con los mapeos dependen en la práctica de los tipos de vocabularios involucrados y su peculiar modo de ordenar el conocimiento. En el caso que nos ocupa, ya hemos referido que el vocabulario de la Biblioteca de la UPM contempla en sus pautas únicamente la utilización de subencabezamientos cronológicos y lo hace de modo casi anecdótico, por lo que podemos asegurar que a efectos de mapeo y como vocabulario fuente estamos ante

un tesoro. Evidentemente el resto de vocabularios tienen una estructura mucho más compleja, que además de las típicas relaciones asociativas, jerárquicas y de relación, tienen vocabularios compuestos de elementos complejos, entre los cuales figuran los de carácter precoordinado, lo cual dificulta la vinculación por el ajuste semántico que se requiere.

Del análisis de vocabularios sacamos algunas conclusiones de procedimiento que podemos enumerar para la gestión eficaz del mapeo:

1. Cuantitativamente las correspondencias entre el vocabulario UPM y los demás vocabularios (específicamente LCSH) no tiene grandes dificultades para establecerse. Son muy escasos los casos donde el mapeo deba realizarse entre conceptos simples o compuestos BUPM y materias complejas o formadas con precoordinación. Reflejamos algunos ejemplos:

En este caso se establece un mapeo de equivalencia inexacta entre la materia BUPM “Cinematografía médica” y la materia BNF “Cinéma -- Applications médicales”.

Cinematografía médica

~EQ Cinéma -- Applications médicales

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/976> a skos:Concept ;
skos:prefLabel "Cinematografía médica"@es ;
skos:notation "778.5:61"@es ;
skos:exactMatch <http://id.loc.gov/authorities/sh85026027#concept> ;
skos:closedMatch <http://data.bnf.fr/ark:/12148/cb13493919r> .
```

Equivalencia inexacta entre encabezamiento compuesto mapeo con encabezamiento formado por precoordinación.

Arte de la Edad Moderna

~EQ Art, Modern--20th century

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/399> a skos:Concept ;
skos:prefLabel "Arte de la Edad Moderna"@es ;
skos:closeMatch <http://id.loc.gov/authorities/sh85007805#concept> ;
skos:scopeNote "Nota editorial. Nuevo concepto. Fuente LEMBP.
Forma precoordinada propuesta: Arte--Historia-- Edad Moderna.
Fecha creación:10062014."@es ;
skos:narrower <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/400> ,
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/401> ,
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/402> ,
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/403> ,
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/412> .
```

2. Las estructuras de mapeo de la LCSH han sido las utilizadas para establecer el resto de mapeos multilinguaje. Los análisis más específicos en cuanto a semántica de mapeos únicamente se han realizado entre los vocabularios BUPM, BNE (como fuente) y LCSH.
3. La tipología de mapeos de la ISO 25964 -2 utilizada aquí es la de vinculación BUPM-LCSH no recíproca. Por lo tanto el tesoro se constituye como vocabulario fuente (BUPM) y la lista de encabezamientos de materia de la Library of Congress como vocabulario objetivo (estableciendo los mapeos secuenciales con los demás vocabularios).
4. La recomendación ISO para este tipo de mapeo es vincular cada concepto del tesoro con su correspondiente semánticamente. Esta tarea se ha realizado aquí de modo limitado, estableciendo mapeos preferentemente entre los *Top Concepts* y el siguiente nivel jerárquico, profundizando en la asignación de mapeos en ramas de especial interés. En no pocas ocasiones, el mapeo selectivo se establece entre vocabularios por cuestiones de economía de recursos, esta práctica, aunque comprensible, no es adecuada pues ocasiona inconvenientes sobre todo a la hora de añadir nuevos conceptos que se relacionan con otros no mapeados en principio.
5. El ajuste entre conceptos y encabezamientos de materia (entre vocabulario fuente y objeto) se define desde distintos puntos de vista: los ajustes semánticos que nos ofrecen las propiedades de mapeo de SKOS: *exactMatch*, *closedMatch*, *relatedMatch*, *broaderMatch* y *narrowerMatch*, y por otro lado está el ajuste entre conceptos según la complejidad de su estructura: conceptos simples, compuestos, precoordinaados, etc. Podemos encontrarnos ante diferentes situaciones:
 - a. Conceptos simples y encabezamientos simples. En este caso se debe efectuar únicamente el ajuste semántico.
 - b. Conceptos simples y encabezamientos complejos. En estos casos el ajuste semántico puede realizarse directamente, o dividiendo el encabezamiento complejo. En no pocos casos los encabezamientos de materia precoordinaados figuran también como encabezamientos simples tras su división. Es el caso de la LCSH, que describe en sus esquemas, tanto la forma compleja, como las formas simples del encabezamiento complejo tras su división (cuestión que es descrita por MADS).
 - c. Conceptos complejos con encabezamientos simples. Se requiere únicamente valoración del ajuste semántico.
 - d. Conceptos complejos con encabezamientos complejos. Se efectúa valoración semántica del ajuste y cabe también la división del encabezamiento complejo en

ambos vocabularios para permitir un ajuste más específico. (Este ajuste sólo es posible si el mapeo es recíproco)

- e. Conceptos simples y complejos donde el concepto en el vocabulario objetivo no existe a priori. En este caso el concepto puede ser establecido según las normas de combinación precoordinada de encabezamientos y subencabezamientos de dicho vocabulario, o por establecimiento de otras formas complejas admitidas. Aquí la norma supone que el establecimiento de mapeos se realiza de modo bidireccional, desgraciadamente esto no es lo habitual y por ello la creación de un encabezamiento precoordinado, por ejemplo en la LCSH, sólo tendría efectos locales, si no existe un IRI que permita la vinculación de los diferentes conceptos en los distintos vocabularios. Es posible que el enfoque aplicación (para el que verdaderamente está definida la norma) permita la composición automática, pero en realidad si en el vocabulario local no existe el concepto como tal y un IRI que lo identifique, no habrá indización de recursos o posibilidades de búsqueda posibles en el contexto de Linked Data. Ponemos un ejemplo del modelo:

Al establecer el mapeo del concepto Fauna, no se consigue la especificidad requerida en la vinculación con el concepto objetivo Animals. Se establece mapeo de equivalencia con dicho concepto por equivalencia inexacta y mapeo asociativo con el concepto Zoology.

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/754> a skos:Concept ;
    skos:prefLabel "Fauna"@es ;
    skos:topConceptOf <http://academia6.poolparty.biz/Tesaurus-
Materias-BUPM/257> ;
    skos:closeMatch
    <http://id.loc.gov/authorities/sh85005249#concept> ;
    skos:scopeNote "Nota editorial: mapeo de equivalencia compuesta
con encabezamiento nuevo creado por precoordinación en el LCSH.
Fauna EQ Animals--Zoology. La forma Animals no se ajusta
completamente al alcance del concepto Fauna. Se decide crear
encabezamiento complejo por precoordinación con los
encabezamientos Animals--Zoology. Forma alternativa
Animals (Zoology).
Fecha creación: 10072014."@es ;
    skos:notation "591.9"@es ;
    skos:relatedMatch
    <http://id.loc.gov/authorities/sh85149983#concept> .
```

- 6. Terminado el mapeo, y teniendo en cuenta que sólo se trabaja de modo unidireccional, la conversión de los conceptos del tesauro, mediante su expresión como términos de búsqueda, permite utilizar el vocabulario BUPM como instrumento de acceso a las

colecciones de recursos indizados por el resto de vocabularios. Evidentemente esto se consigue a nivel de aplicación.

5.3.2.3 *Supuestos de mapeo de equivalencia compuesta*

Como se ha apuntado, las propiedades SKOS utilizadas en este tipo de mapeos son *skos:exactMatch* y, *skos:closedMatch*, con la diferencia de que la primera es una propiedad transitiva, es decir, establecida en la cabecera de la jerarquía no es necesario repetirla en cada nivel y que supone un ajuste perfecto entre la semántica de los conceptos mapeados. En el ámbito de la norma ISO 25964 – 2, estos grados se expresan mediante los símbolos = y ~, el primero para la equivalencia exacta (reversible) y el segundo para la inexacta. Como se ve, su transposición a SKOS no presenta dificultades:

Agricultura de regadío

=EQ Irrigation farming

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/625> a skos:Concept ;
    skos:prefLabel "Agricultura de regadío"@es ;
    skos:notation "631"@es ;
    skos:exactMatch
    <http://id.loc.gov/authorities/sh85068298#concept> .

<http://id.loc.gov/authorities/sh85068298#concept> a skos:Concept ;
    skos:prefLabel "Irrigation farming"@en .
```

Estructuras (Construcción)

~EQ Structural engineering

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/632> a skos:Concept ;
    skos:prefLabel "Estructuras (Construcción)"@es ;
    skos:closeMatch <http://id.loc.gov/authorities/sh85129198#concept> ;
    skos:narrowMatch <http://id.loc.gov/authorities/sh85129216#concept> .

<http://id.loc.gov/authorities/sh85129198#concept> a skos:Concept ;
    skos:prefLabel "Structural engineering"@en .

<http://id.loc.gov/authorities/sh85129216#concept> a skos:Concept ;
    skos:prefLabel "Structural analysis (Engineering)"@en .
```

En cualquiera de los casos expuestos más abajo, cuando la equivalencia es compuesta, es muy improbable que los significados sean absolutamente equivalentes, podríamos decir que toda equivalencia compuesta es inexacta. Existe un supuesto no contemplado en este trabajo y regulado en la norma: el mapeo equivalente parcial. Se trata de relaciones jerárquicas donde el

alcance del concepto más genérico no se diferencia mucho del alcance del más específico; en muchos caso la distinción entre utilización de equivalencia parcial o no depende de si en el vocabulario se establecen estructuras jerárquicas o no.

Ya se explicó en el capítulo anterior el sistema de relaciones de mapeo que dispone la norma ISO 25964 -2, el cual se orienta más a la estructuración de tesauros en el contexto de las aplicaciones que en un contexto distribuido en red. Es por ello que algunas categorías de relaciones de equivalencia no puedan ser representadas en SKOS de un modo totalmente eficaz. En los casos de equivalencia compuesta cumulativa, como el siguiente, la relación se establece mediante la unión de dos conceptos en el vocabulario objetivo: el primero tiene un alcance más genérico que la expresión compleja del vocabulario fuente, mientras el segundo concepto especifica al primero y se modela como relacionado. Como se puede observar, los dos conceptos son absolutamente independientes (Este caso se diferencia del anterior apuntado, ejemplo Fauna EQ Animals - - Zoology en que no se pretende la creación de un encabezamiento complejo en el vocabulario objetivo):

Denominación de origen Jerez

EQ Marks of origin | Jerez de los Caballeros Region (Spain)

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/580> a skos:Concept ;
    skos:prefLabel "Denominación de origen Jerez"@es ;
    skos:broadMatch
    <http://id.loc.gov/authorities/sh85081371#concept> ;
    skos:relatedMatch
    <http://id.loc.gov/authorities/sh90004722#concept> .

<http://id.loc.gov/authorities/sh85081371#concept> a skos:Concept ;
    skos:prefLabel "Marks of origin"@en.

<http://id.loc.gov/authorities/sh90004722#concept> a skos:Cocncept ;
    skos:prefLabel "Jerez de los Caballeros Region (Spain)"@en .
```

Ya se refirió en otras partes de este trabajo la utilidad de introducir los operadores booleanos en la gramática de SKOS como posible revisión de su vocabulario.

Otro supuesto de mapeo es el de equivalencia compuesta con intersección de los alcances de los conceptos del vocabulario, los dos conceptos involucrados en el vocabulario objetivo tienen sus espacios semánticos con un cierto nivel de relación:

Matemáticas financieras EQ Mathematics + Business

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/582>a skos:Concept ;
    skos:prefLabel "Matemáticas financieras"@es ;
    skos:broadMatch <http://id.loc.gov/authorities/sh85082139#concept> ,
    <http://id.loc.gov/authorities/sh85018260#concept> ;

<http://id.loc.gov/authorities/sh85082139#concept> a skos:Concept ;
    skos:prefLabel "Mathematics"@en .

<http://id.loc.gov/authorities/sh85018260#concept> a skos:Cocncept ;
    skos:prefLabel "Business"@en .
```

Otros casos de equivalencia suponen escoger encabezamientos complejos del vocabulario objetivo, frente a conceptos simples en el vocabulario fuente. En este caso, el concepto “Gravimetría” no aparece como tal en la LCSH, pero si se encuentra la combinación *Gravity + Measurement* unidos como elemento precoordinado. En este caso estamos ante un supuesto simple de equivalencia, pues el concepto obtenido mediante precoordinación se considera un único concepto. Es habitual que uno de los conceptos de la cadena precoordinada sea concepto genérico respecto al concepto del vocabulario fuente, como ya hemos apuntado.

Gravimetría EQ Gravity -- Measurement

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/589> a skos:Concept ;
    skos:prefLabel "Gravimetría"@es ;
    skos:notation "543.21"@es ;
    skos:closeMatch <http://id.loc.gov/authorities/sh85056564#concept> ;
    skos:broadMatch <http://id.loc.gov/authorities/sh85056563#concept> .

<http://id.loc.gov/authorities/sh85082564#concept> a skos:Concept ;
    skos:prefLabel "Gravity--Measurement"@es .

<http://id.loc.gov/authorities/sh85018563#concept> a skos:Cocncept ;
    skos:prefLabel "Gravity"@es .
```

En este último caso la equivalencia se establece entre un concepto compuesto del vocabulario fuente y un encabezamiento complejo compuesto por un encabezamiento + un delimitador + un subencabezamiento en el vocabulario objetivo (International Organization for Standardization (ISO), 2012).

Planificación rural

EQ Land use, Rural - - Planning

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/605> a skos:Concept ;
    skos:prefLabel "Planificación rural"@es ;
    skos:closeMatch <http://id.loc.gov/authorities/sh2009128698#concept> ;
    skos:notation "711.3"@es ;
    skos:altLabel "Ordenación rural"@es .

<http://id.loc.gov/authorities/sh85082698#concept> a skos:Concept ;
    skos:prefLabel "Land use, Rural--Planning"@en .
```

5.3.2.4 Relaciones jerárquicas

Las relaciones jerárquicas en mapeos se establecen en similares condiciones a las relaciones internas. Así es para los mapeos jerárquicos genéricos y de instancia, mientras que la norma ISO 25964-2 refiere que la utilización del tipo parte-todo, puede ser conflictivo en este contexto.

Del análisis efectuado en los vocabularios aquí utilizados se deduce que no es un tipo de relación muy utilizado, siendo más frecuentes las relaciones de equivalencia y en menor medida las asociativas. Tampoco refiere la norma en qué circunstancias se debe utilizar las relaciones jerárquicas en mapeos, pues en muchos casos es posible que no sea necesaria pues ya ha sido establecida en niveles superiores de la jerarquía.

Para el tesauro de materias BUPM, las relaciones jerárquicas en mapeos sólo se utilizarán en el caso de que dicha vinculación mejore semánticamente el alcance del concepto vinculado. Por ejemplo, en el siguiente caso, el concepto “Agrimensura” no tiene un equivalente definido en LCSH ni en los demás vocabularios. Se decide establecer un mapeo jerárquico de tipo genérico con *Surveys* que aunque en una traducción literal no define el nivel genérico de “Agrimensura” si tiene una semántica que enmarca el concepto anterior desde el punto de vista del uso que le da LCSH y por tanto a las indizaciones efectuadas con dicho concepto. (Podría haberse establecido, conectando con el apartado anterior una relación de equivalencia inexacta):

Agrimensura **BMG Surveys**

```
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/621> a skos:Concept ;
    skos:prefLabel "Agrimensura"@es ;
    skos:broadMatch <http://id.loc.gov/authorities/sh99001768#concept> ;
    skos:closeMatch <http://data.bnf.fr/ark:/12148/cb12306118r> ;
    skos:closeMatch dbpedia:Land_surveys ;
    skos:notation "528.4"@es .

<http://id.loc.gov/authorities/sh99001768#concept> a skos:Concept ;
    skos:prefLabel "Surveys"@en .

<http://data.bnf.fr/ark:/12148/cb12306118r> a skos:Concept ;
    skos:prefLabel "Leves"@fr .

dbpedia:Land_surveys a skos:Concept .
```

La notación que dispone la norma ISO para el marcado de mapeos es la siguiente:

BM (*BroaderMapping*) y NM (*NarrowerMapping*). Se pueden establecer diferenciaciones de tipo mediante la adición de G (genérica), I (de instancia) ó P (parte-todo): BMG, BMI, BMP; NMG, NMI, NMP.

5.3.2.5 Las relaciones asociativas

Las relaciones asociativas en mapeos siguen los mismos parámetros que las que se establecen en el propio tesauro. La norma ISO 25964-2 presenta el indicador RM (*RelatedMapping*) para establecerlos. Al igual que con los tesauros, la distancia entre relaciones asociativas y equivalentes inexactas no es amplia y permite al editor del tesauro escoger entre una u otra dependiendo del alcance de los conceptos implicados o cuestiones específicas de los vocabularios como su ámbito o las costumbres de búsqueda de los usuarios. En el contexto de este trabajo las

relaciones de equivalencia inexacta se establecerán entre los conceptos mapeados si existe un ajuste importante de su significado mientras que los mapeos relacionados se utilizarán para conectar campos semánticos suficientemente diferenciados, pero que conviene relacionar para describir conexiones puntuales que ayuden en la indización o en la búsqueda. Se modela algún ejemplo de mapeo asociativo:

Transporte aéreo

RM Transportation engineering

```
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/661> a skos:Concept ;
    skos:prefLabel "Transporte aéreo"@es ;
    skos:closeMatch <http://id.loc.gov/authorities/sh85001360#concept> ;
    skos:notation "656.7"@es ;
    skos:relatedMatch <http://id.loc.gov/authorities/sh85137069#concept> .
```

```
<http://id.loc.gov/authorities/sh85001360#concept> a skos:Concept ;
    skos:prefLabel "Aeronautics, Commercial"@en .
```

```
<http://id.loc.gov/authorities/sh85137069#concept> a skos:Concept ;
    skos:prefLabel "Transportation engineering"@en .
```

(International Organization for Standardization (ISO), 2012)

5.3.2.6 Documentación de mapeos

La norma ISO 25964-2 especifica la necesidad de documentar la información que afecte a los mapeos. Ya se ha hablado aquí de la importancia de la descripción de datos y datasets. Los mapeos no son una excepción y en lo que concierne a este proyecto se declararán metadatos descriptivos Dublin Core en el fichero VoID bajo la categoría de cada *void:linkset*. En esta información general se describirán los períodos temporales preceptivos, el tipo de generación de los mapeos, un etiquetado identificativo, información administrativa (editor, mantenedor, etc), información relevante para el acceso, y cualquier cuestión afectante al copyright.

La información general del mapeo de conceptos aparece en el gestor y las especificidades en el sistema de anotaciones. Si del desarrollo de los mapeos se definieran agrupaciones de los mismos, se describirán esos conjuntos de mapeos. También se debe describir cualquier información que restrinja el ámbito y las posibilidades de mapeo. Si la agrupación se verifica mediante clúster debe declararse la información de identificación, organización, administración y copyright.

A continuación se expone la información suministrada para la descripción del mapeo en VOID (publicado en CKAN) con un nivel de granularidad básico:

.....

```
<http://upm.es/biblioteca/kos/sh/bibupm-LCSH-01> a void:Linkset;
    dcterms:type "mapping" , "one-way mapping";
    dcterms:creator "R.A.A.";
    dcterms:iissued "08012014";
    dcterms:valid "2016";
    dcterms:description "Set de mapeos establecidos entre LEM BUPM y LCSH";
    dcterms:publisher "R.A.A.";
    dcterms:VocabularyEncodingScheme <http://www.w3.org/2004/02/skos/core/>;
    dcterms:provenance "Datos en silos BUPM (Oracle)- Publicación LEM en LOD
http://.....";
    dcterms:license "CC BY 4.0";
    dcterms:instructionalMethod "ISO 25964-2.";
    dcterms:isVersionof <http://upm.es/biblioteca/kos/sh/bibupm-LCSH-00>
    dcterms:hasVersion <http://upm.es/biblioteca/kos/sh/bibupm-LCSH-02>
```

.....

(Cyganiak, Zhao, Alexander, & Hausenblas, 2010; International Organization for Standardization (ISO), 2012)

5.3.2.7 Mantenimiento y preservación de mapeos

La norma ISO 25964-2 establece la necesidad de proveer sistemas de almacenado para los mapeos. En un proyecto como éste la gestión del almacenamiento se efectúa a través de los *triple stores* de Semantic Web y de las copias de seguridad y almacenamiento en *data dumps* establecidos como estrategia de preservación. Para la compatibilidad futura, la propia norma favorece e indica la utilización de SKOS como estándar para la estructuración de los set de mapeos.

El problema fundamental aquí es la sostenibilidad de los vínculos entre conceptos, cuestión para la que el versionado de los mapeos es fundamental y para la que hay que tener en cuenta los cambios posibles en los datos. El control de versiones de mapeo puede instrumentarse en ficheros de metadatos *Provenance*.

Tabla 7 Gestión de actualizaciones en los mapeos. ISO 25964 (2012).

| Cambios en el vocabulario fuente | Consecuencias en los mapeos |
|---|-----------------------------|
| Nuevo concepto | Nuevo proceso de mapeo |
| Concepto eliminado | Eliminar el mapeo |
| Variación del concepto (variación a concepto complejo, división, etc) | Rehacer los mapeos |
| Cambios en el alcance de los conceptos | Corregir mapeos |

La norma también refiere la necesidad de establecer una política de comunicación de cambios efectiva y de replantear los sistemas de indización y de búsqueda que dependan de los vocabularios.

El problema de la preservación debe contemplarse como nuclear en este tipo de proyectos. En el capítulo siguiente se establecen métodos válidos para la preservación de datos y datasets, pero los mapeos requieren alguna precisión especial además de la referida de la identificación de las versiones. Generar sistemas automáticos que comprueben la válida conexión entre conceptos debe ser un objetivo de desarrollo prioritario, si además cuentan con la característica de corrección automática mejor. La calidad y la propia credibilidad del sistema están en juego.

(International Organization for Standardization (ISO), 2012)

5.4 SKOSIFICACION

Aunque teóricamente es posible la estructuración de vocabularios controlados con la combinación de RDF, RDFS y OWL, SKOS se ha convertido en el estándar para el modelado de dichos vocabularios, tarea para la que ha sido diseñado específicamente. SKOS, como ya se dijo, no permite una descripción formal ni exhaustiva de un vocabulario. En el terreno de lo ideal se debería implementar una ontología basada en alguna de las variantes de OWL, lo que aseguraría una perfecta disposición del vocabulario incluso en su “exportación” a la web de datos, al precio de disponer de una gran cantidad de recursos para ello, tanto en la creación como el mantenimiento. Lo que verdaderamente ofrece SKOS es la posibilidad de hacer efectivo el modelado semántico de nuestro vocabulario de modo factible, con un gasto moderado de recursos y con un nivel de complejidad asumible para la mayoría de los editores.

No todos los vocabularios tienen el mismo acomodo bajo SKOS, por ejemplo, no es muy adecuado para listas de autoridades personales, pero si encaja mejor con las estructuras de tesauros y encabezamientos de materia.

Modelar en SKOS para la Web de datos supone adaptarnos al modelo de tesauros conceptuales. En este proyecto se sigue esa pauta y se ha verificado la distinción entre conceptos y sus etiquetas (términos) que los representan. También y como consideraciones generales hemos de evaluar antes de su utilización si nuestro vocabulario tiene en cantidades importantes, nombres de personas, de lugares, de instituciones o distinciones de género y número, en estos casos SKOS puede no ser suficiente.

Del análisis derivado del vocabulario aquí desarrollado se deduce la viabilidad de utilizar SKOS para su modelado. El modelo SKOS permite representar con suficiencia el conocimiento contenido en el vocabulario BUPM. Anteriormente nos hemos referido a sus carencias representativas, sobre todo con expresiones conceptuales complejas o con mapeos de equivalencia complicados. Ahora haremos frente a las limitaciones del gestor, cuya expresividad

no ha alcanzado los niveles previstos en un principio. En primer lugar se exponen aquí las propuestas de modelado SKOS a efectuar manualmente, las cuales serán incorporadas al fichero completo del tesoro alojado en The DataHub. (Para distinguir el modelado manual se ha formateado en **negrita**):

1. A nivel conceptual SKOS permite una vertebración del vocabulario suficiente: generación de esquemas generales (*ConceptScheme*) y su contenido (*hasTopConcept*), identificación de conceptos (*Concepts*) y las adscripción de estos a un esquema determinado (*inScheme*) y su categorización como concepto de cabecera (*topConceptOf*). Nuestro gestor de tesauros no permite utilizar la propiedad *inScheme* lo que supone que la consulta de aplicaciones pueda tener dificultades para vincular un concepto con su esquema.

```
<http://academia6.poolparty.biz/Tesoro-Materias-BUPM/623> a skos:Concept ;
    skos:prefLabel "Agricultura alternativa"@es ;
    skos:notation "631.147"@es ;
    skos:narrower
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/628> ;
    skos:inScheme
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/257>
    skos:exactMatch
    <http://id.loc.gov/authorities/sh87005122#concept> ;
    skos:relatedMatch
    <http://id.loc.gov/authorities/sh85002482#concept> .
```

2. Las colecciones de conceptos también son modelables con SKOS. En este proyecto se presentan dos colecciones para agrupar dos clases de conceptos muy numerosos en el Área 5, los programas de ordenador y los lenguajes de programación y por otro lado un grupo de subencabezamientos cronológicos. El software gestor no modela con las clases *Collection* o *member*, en vez de ello asigna una etiqueta de nodo a los conceptos, bajo un orden determinado por una *query* específica de SPARQL (Semantic Web Company GmbH, 2012) y sin definir ninguna colección como clase. Esta configuración no permite la identificación de colecciones identificadas semánticamente con SKOS. En nuestro tesoro las colecciones modelan *arrays* que agrupan conceptos semánticamente acumulables y encabezados por una etiqueta de nodo. En el ejemplo siguiente la etiqueta de nodo es <Lenguajes de programación>, se declara la colección como clase *skos:Collection* y se identifican sus componentes mediante la propiedad *skos:member*. Se identifica en rojo el modelado “sui generis” del programa y en **negrita** el modelado manual de la colección.

```
<http://academia6.poolparty.biz/Tesoro-Materias-BUPM/283> a skos:Concept ;
    skos:prefLabel "Lenguajes de programación orientados a objetos"@es ;
    skos:notation "004.432"@es ;
    skos:altLabel "Lenguajes OO (Lenguajes de programación)"@es , "Lenguajes
    OOP (Lenguajes de programación)"@es ,
```

```

    "Lenguajes orientados a objetos (Lenguajes de programación)"@es ;
    skos:narrower
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/284> ,
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/285> ,
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/345> .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/285>a skos:Concept ;
    skos:prefLabel "C# (Lenguaje de programación)"@es ;
    skos:notation "004.432"@es ;
    skos:altLabel "C-Sharp (Lenguaje de programación)"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/283> ;
    <http://schema.semantic-web.at/sparql/1.0/list/orderedBy>
    _:node18s157hpjx5 .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/284> a skos:Concept ;
    skos:prefLabel "JavaScript (Lenguaje de programación)"@es ; skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/283> .
    skos:notation "004.43"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/283> ;
    skos:exactMatch
    <http://id.loc.gov/authorities/sh96004880#concept> ;
    <http://schema.semantic-web.at/sparql/1.0/list/orderedBy>
    _:node18s157hpjx6 .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/345> a skos:Concept ;
    skos:prefLabel "Visual dBASE (Lenguaje de programación)"@es ;
    skos:notation "004.65"@es ;
    skos:altLabel "Borland Visual dBASE (Lenguaje de programación)"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/283> ;
    <http://schema.semantic-web.at/sparql/1.0/list/orderedBy>
    _:node18s157hpjx16 .

_:node18s157hpjx1 a skos:OrderedCollection;
    skos:prefLabel "<Lenguajes de Programación>"@es;
    skos:member
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/284> ,
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/285> ,
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/345> .

```

3. En cuanto a la descripción léxica, SKOS nos ofrece la posibilidad de identificar el término preferido (*prefLabel*), alternativo (*altLabel*) u oculto (*hiddenLabel*), incluso nos permite establecer relaciones entre etiquetas utilizando la extensión SKOS-XL. Estas posibilidades se ajustan a las necesidades de modelado del tesoro de materias, permitiendo la vinculación de diferentes etiquetas como ocurre en las expresiones precoordinaadas o enumerando sus componentes. El gestor incluye todas las etiquetas léxicas de SKOS, pero

no así la extensión SKOS-XL, muy útil para los casos en los que se requiera la expresión de términos diferenciados por número o género. En su conjunto son suficientes para expresar las relaciones de equivalencia, pero no pueden recoger la mayor riqueza definida en la norma ISO 25964 -1, la cual hace, como vimos, un uso extensivo y complejo de la equivalencia. Para superar estos inconvenientes se ha completado manualmente el modelado ofrecido por defecto por el gestor, para una rama jerárquica, cuyo *TopConcept* es “Ciencias sociales” Se incluye la descripción de conceptos complejos precoordinaados presentes en el tesauo de materias, utilizando agregadamente el vocabulario MADS y la extensión SKOS-XL.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#> .

<http://academia6.poolparty.biz/Tesauo-Materias-BUPM/825> a skos:Concept ;
    skos:prefLabel "Ciencias Sociales"@es ;
    skos:topConceptOf
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/261> ;
    skos:notation "3"@es ;
    skos:narrower
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/729> ,
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/826> .

<http://academia6.poolparty.biz/Tesauo-Materias-BUPM/729> a skos:Concept ;
    skos:prefLabel "Estadística"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/825> ;
    skos:closeMatch <http://id.loc.gov/authorities/sh99001414#concept> ;
    skos:notation "311"@es .

<http://academia6.poolparty.biz/Tesauo-Materias-BUPM/829> a skos:Concept ;
    skos:prefLabel "Historia contemporánea"@es ;
    skos:narrower
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/820> ;
    skos:broader
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/826> ;
    skos:closeMatch <http://id.loc.gov/authorities/sh85061236#concept> .

#Conceptos utilizados para la construcción del concepto complejo
precoordinado#

<http://academia6.poolparty.biz/Tesauo-Materias-BUPM/826> a skos:Concept ;
    a madsrdf:Topic, madsrdf:Authority ;
    madsrdf:authoritativeLabel "Historia"@es ;
    skos:prefLabel "Historia"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/825> ;
    skos:narrower
    <http://academia6.poolparty.biz/Tesauo-Materias-BUPM/829> ;
```

```

skos:closeMatch <http://id.loc.gov/authorities/sh85061212#concept> .

#Se incluye la descripción MADS de los conceptos geográficos "Europa y España"
usados como subencabezamientos geográficos y el concepto temporal "S. XIX"
utilizado como subencabezamiento temporal. Su descripción estaría en la
secuencia jerárquica correspondiente, pero se incluye aquí para una mejor
lectura del modelado. Para una correcta especificación se debería crear un
esquema contenedor de subencabezamientos.#

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/812> a skos:Concept ;
    a madsrdf:Geographic, madsrdf:Authority ;
    madsrdf:authoritativeLabel "España"@es ;
    skos:prefLabel "España"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/811> ;
    skos:exactMatch <http://id.loc.gov/authorities/sh85126198#concept> .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/813> a skos:Concept ;
    a madsrdf:Geographic, madsrdf:Authority ;
    madsrdf:authoritativeLabel "Europa"@es ;
    skos:prefLabel "Europa"@es ;
    skos:exactMatch <http://id.loc.gov/authorities/sh85045631#concept> ;
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/811> .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/815> a skos:Concept ;
    a madsrdf:Temporal, madsrdf:Authority ;
    madsrdf:authoritativeLabel "S. XIX"@es ;
    skos:prefLabel "S. XIX"@es ;
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/814> ;
    skos:closeMatch
    <http://id.loc.gov/authorities/sh2002012475#concept> ;
    skos:exactMatch <http://id.loc.gov/authorities/sh85091984#concept> .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/820> a skos:Concept ;
    skos:prefLabel "Europa -- Historia -- S. XIX"@es ;
#La Library of Congress genera una cadena de conceptos en secuencia que
identifican el concepto complejo coordinado mediante SKOS-XL#
    skosxl:altLabel [
        a skosxl:Label ;
        skosxl:literalForm "Europa, Historia, S. XIX"@es ;
    ]
    skos:broader
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/829> ;
    skos:narrower
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/819> ;
    skos:inScheme
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/261> ;
    skos:closeMatch
    <http://id.loc.gov/authorities/sh85126088#concept> ;
    a madsrdf:ComplexSubject, madsrdf:Authority ;
    madsrdf:authoritativeLabel "Europa -- Historia -- S. XIX"@es ;
    madsrdf:isMemberOfMADSScheme
    <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/> ;
    madsrdf:componentList (
        <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/813>

```

```

<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/826>
<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/815>
) .

<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/819> a skos:Concept ;
    skos:prefLabel
        "España -- Historia -- Guerra de la Independencia, 1808-1814"@es ;
    skosxl:altLabel [
        a skosxl:Label ;
        skosxl:literalForm
            "España, Historia, Guerra de la Independencia, 1808-1814"@es ;
    skos:broader
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/829> ;
    skos:narrower
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/830> ;
    skos:closeMatch <http://id.loc.gov/authorities/sh85126089#concept>
    skos:inScheme
        http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/261 ;
    a madsrdf:ComplexSubject , madsrdf:Authority ;
    madsrdf:authoritativeLabel
        "España -- Historia -- Guerra de la Independencia, 1808-1814"@es ;
    madsrdf:isMemberOfMADSScheme
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/> ;
    madsrdf:componentList (
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/812>
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/826>
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/830>
    ) .

<http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/830> a skos:Concept ;
    a madsrdf:Topic, madsrdf:Authority ;
    madsrdf:authoritativeLabel
        "Guerra de la Independencia, 1808-1814"@es ;
    skos:prefLabel "Guerra de la Independencia, 1808-1814"@es ;
    skos:broader
        <http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/829> ;
    skos:closeMatch <http://id.loc.gov/authorities/sh85099422#concept> .

```

4. Las posibilidades de anotación de SKOS (*scopeNote*, *changeNote*, *definition*, *editorialNote*, *etc.*) permiten una extensa documentación de todos los aspectos del tesauro. Estas posibilidades son suficientes para las anotaciones del tesauro de materias, con especial referencia a las notas editoriales necesarias para explicar la asignación de conceptos, sus fuentes y su estado de validez. Las notas de alcance, muy útiles para precisar campos semánticos, no han sido muy utilizadas aquí, debido a que es determinación semántica ya estaba efectuada fundamentalmente en el vocabulario original. El gestor sólo contempla la inclusión de la nota de alcance (lugar elegido para definir los contenidos de las notas editoriales y otras especificaciones).

5. También se ha incluido el atributo *@xml:lang*, aunque no estamos ante un tesoro multilingüe a nivel interno. Con ello se prevé la posible ampliación de idiomas en el vocabulario de materias.
6. Las relaciones que ofrece SKOS se estiman suficientes para el desarrollo de este tesoro. Tanto las relación concepto genérico-específico como la de concepto relacionado. No recoge el gestor las relaciones transitivas (*broaderTransitive* y *narrowerTransitive*) propiedades que hubieran sido útiles para afinar una vinculación más específica entre conceptos en relación jerárquica más compleja. Paradigmático es un caso especialmente importante para “informar” a las aplicaciones de la existencia de una relación genérico-específico entre Teoría de la computación y Automatas celulares, Autómatas finitos y Autómatas conscientes.

```

<http://academia6.poolparty.biz/Tesoro-Materias-BUPM/352>
  a skos:Concept ;
  skos:prefLabel "Teoría de la computación"@es ;
  skos:topConceptOf
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/260> ;
  skos:notation "004"@es ;
  skos:altLabel "Computación, Teoría de la"@es ;
  skos:narrower
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/353> ,
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/361> .

<http://academia6.poolparty.biz/Tesoro-Materias-BUPM/353> a skos:Concept ;
  skos:prefLabel "Teoría de autómatas"@es ;
  skos:notation "004"@es ;
  skos:altLabel "Autómatas, Teoría de"@es ;
  skos:broader
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/352> ;
  skos:narrower
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/354> ,
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/355> ,
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/356> .

<http://academia6.poolparty.biz/Tesoro-Materias-BUPM/354> a skos:Concept ;
  skos:prefLabel "Autómatas celulares"@es ;
  skos:notation "004.383.8"@es ;
  skos:altLabel "Ordenadores de circuito iterativo"@es ;
  skos:broader
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/353> ;
  skos:broaderTransitive
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/352> .

<http://academia6.poolparty.biz/Tesoro-Materias-BUPM/356> a skos:Concept ;
  skos:prefLabel "Autómatas finitos"@es ;
  skos:notation "007.52"@es ;
  skos:altLabel "Máquinas de Mealy y Moore"@es .
  skos:broader
    <http://academia6.poolparty.biz/Tesoro-Materias-BUPM/353> ;

```

```
skos:broaderTransitive  
<http://academia6.poolparty.biz/Tesauro-Materias-BUPM/352> .
```

```
<http://academia6.poolparty.biz/Tesauro-Materias-BUPM/355> a skos:Concept ;  
  skos:prefLabel "Autómatas conscientes"@es ;  
  skos:notation "004.89"@es ;  
  skos:broader  
    <http://academia6.poolparty.biz/Tesauro-Materias-BUPM/353> ;  
  skos:broaderTransitive  
    <http://academia6.poolparty.biz/Tesauro-Materias-BUPM/352> .
```

7. Una de las áreas más desarrolladas en SKOS es la del mapeo, donde las propiedades disponibles permiten reflejar de modo bastante fidedigno el conjunto de relaciones posibles entre conceptos en diferentes vocabularios. Estas propiedades son suficientes para la representatividad del vocabulario y además, están representadas íntegramente en el sistema gestor. La construcción del modelado ya ha sido especificada en epígrafes anteriores. La adición de mapeos con el resto de vocabularios se ha agregado manualmente.

6 PUBLICACIÓN DE LA LISTA DE ENCABEZAMIENTOS DE MATERIA DE LA BUPM EN LINKED OPEN DATA

6.1 PUBLICACIÓN DE DATASETS

Quizás el paso más importante a la hora de transformar cualquier vocabulario en un recurso Linked Data es la publicación. Publicar supone exponer nuestro trabajo a los ojos de la comunidad, y ofrecerlo siguiendo ciertos estándares de publicación que permitan su total aprovechamiento y reutilización. La publicación en sedes específicas mejora la visibilidad Web y permite la interacción con los datos y metadatos a través de interfaces de consulta y exploración, además facilita la disposición de los datos a través de su descarga en diferentes formatos y cualquier otra utilidad que facilite su uso.

Publicar también significa identificar, es decir permitir la consulta de metadatos significativos que ayuden a verificar en qué consiste exactamente el dataset que se publica, su procedencia y el alcance de las licencias que permitan la reutilización. Hasta aquí lo que podríamos llamar efectos externos de la publicación. Frente a ellos se define lo que podría denominarse visibilidad interna, es decir las cuestiones estructurales que influyen decisivamente en la publicación:

1. El ofrecimiento de un producto multilingüe que amplíe el rango de utilización.
2. La gestión eficaz de los enlaces en mapeos que por sí mismos constituyen puentes hacia otros vocabularios.
3. La planificación y el mantenimiento de una política de calidad de toda la estructura de datos, definiendo la actualización, el mantenimiento y la preservación.

Que nuestro producto sea de calidad puede tener importantes consecuencias para la visibilidad; un vocabulario determinado y en concreto uno de materias, si es reconocido por la calidad (que no cantidad) de los datos y estructura que ofrece, puede “atraer” enlaces de otras plataformas Linked Data, lo que proveerá, vía interoperabilidad, mayor visibilidad a nuestro producto. La calidad desde un punto de vista amplio, es la mejor manera de generar enlaces productivos, los IRIs son una herramienta que posibilita la conexión y los mapeos una estructura de vinculación, la calidad supone el colofón del sistema, sin ella las autopistas de datos estarán vacías.

6.1.1 PLATAFORMAS DE PUBLICACIÓN. POOLPARTY

Las consideraciones locales de publicación de este proyecto se han visto limitadas por las características de los recursos disponibles para llevarlo a cabo. La herramienta elegida para la gestión del tesoro (PoolParty, de Semantic Web) ofrece interesantes propuestas de publicación, pero se ve limitada por sus características en la versión académica.

Lo ideal es contar con una infraestructura suficiente de publicación, que nos permita tener independencia respecto a soluciones propietarias, a la vez que nos otorga el control absoluto de nuestros datos. Pero una de las opciones que se tienen que tener en cuenta en la fase de publicación es que no siempre tenemos posibilidades de hacerlo por nuestros medios y es preciso adaptarse a las condiciones existentes. En cualquier caso PoolParty, en su versión estándar, sí permite un menor nivel de dependencia, pues ofrece a sus clientes potentes soluciones de publicación de datos totalmente integradas en sus propios sistemas y servidores, lo que permite ofrecer un producto final totalmente personalizado.

Aun así, no es desdeñable el conjunto de características que ofrece PoolParty versión académica para la publicación de datasets:

1. Almacenamiento de datos en un triple store.
2. Publicación de la interfaz humana mediante wiki pages.
3. Múltiples formatos de descarga de ficheros de datos (RDF/XML, Turtle, JSON-LD, TRIG, etc).
4. SPARQL Endpoint para la consulta de datos.
5. Potentes utilidades de actualización del vocabulario.
6. Interfaz de visualización gráfica de los datos.
7. Sistemas de preservación mediante almacenamiento de copias de seguridad en la nube.
8. Reportes de estructura y calidad del dataset.
9. Control configurable de la calidad de datos y estructuras de modelado.
10. Modelado con metadatos descriptivos de modo semiautomático.

Las limitaciones de la plataforma gestora de tesauros, han obligado a reformar algunos aspectos nucleares del proyecto. Por ejemplo, la imposibilidad en la versión no comercial de configurar un dominio propio y tener por ello control absoluto sobre la construcción de los IRIs, o la poca flexibilidad del sistema de mapeos que únicamente permite la vinculación automática con algunos vocabularios, aunque los que ofrece por defecto son de gran importancia: Geonames, Dbpedia, LCSH, etc. (Semantic Web Company GmbH, 2014)

The **PoolParty** approach for efficient knowledge modeling

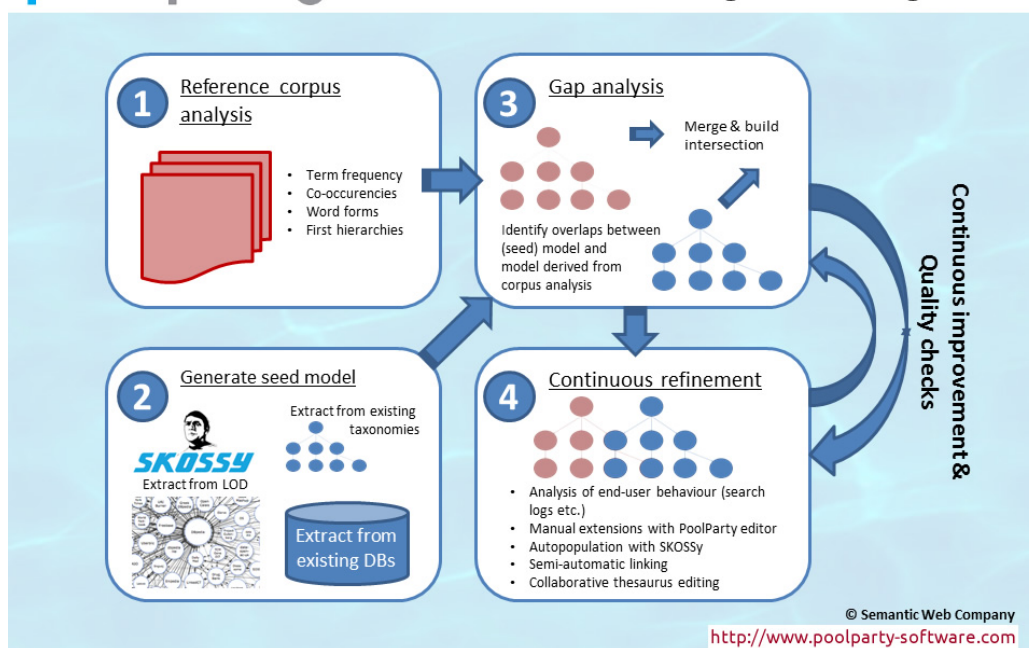


Figura 24 Proceso de datos del gestor de tesauros Poolparty. Fuente: Semantic Web Company, 2014

Su modelo de gestión y creación de vocabularios es bastante completo como se puede ver en el gráfico superior. Se permiten extracciones automáticas de conceptos desde un corpus de documentos; la generación en paralelo de estructuras jerárquicas desde vocabularios ya establecidos o mediante construcción manual; el ajuste de ambos métodos de ingesta generando una versión conjunta; y finalmente, la concreción del vocabulario como proceso continuo y enriquecido por las aportaciones de usuarios finales, las adiciones de profesionales, la generación automática de conceptos con SLOSSY y los mapeos semiautomáticos con los vocabularios antes referidos.

PoolParty ofrece pues, incluso en su versión básica, una estructura de publicación de cierto nivel, que permite una presentación satisfactoria de resultados. La posibilidad de pago en combinación con servidores propios como infraestructura de publicación, ofrece una plataforma de gran nivel como lo demuestra la utilización por grandes entidades del mundo cultural, como la British Library, SMB, o importantes organizaciones como la Comunidad Europea. (Semantic Web Company GmbH, 2014).



Figura 25 Detalle de registro del tesauro de materias BUPM. Fuente: (Ávila, 2014b).

Las publicaciones Linked Data están poniendo el acento en ofrecer posibilidades enriquecidas de visualización. Las mejores propuestas de datos vinculados ofrecen una interfaz de visualización que ayuda de modo importante a extraer la información de los datos, a veces demasiado oculta en los códigos y serializaciones. La interfaz visual de PoolParty, no por sencilla deja de tener una cierta versatilidad a la hora de obtener una visión global de la estructura del tesauro. En la figura inferior podemos observar como las formas circulares y las porciones significativas de colores expresan una particular situación semántica de los términos. Aunque su interpretación no es muy obvia a primera vista, el estudio de jerarquías completas se facilita con un mínimo entendimiento de la metáfora.

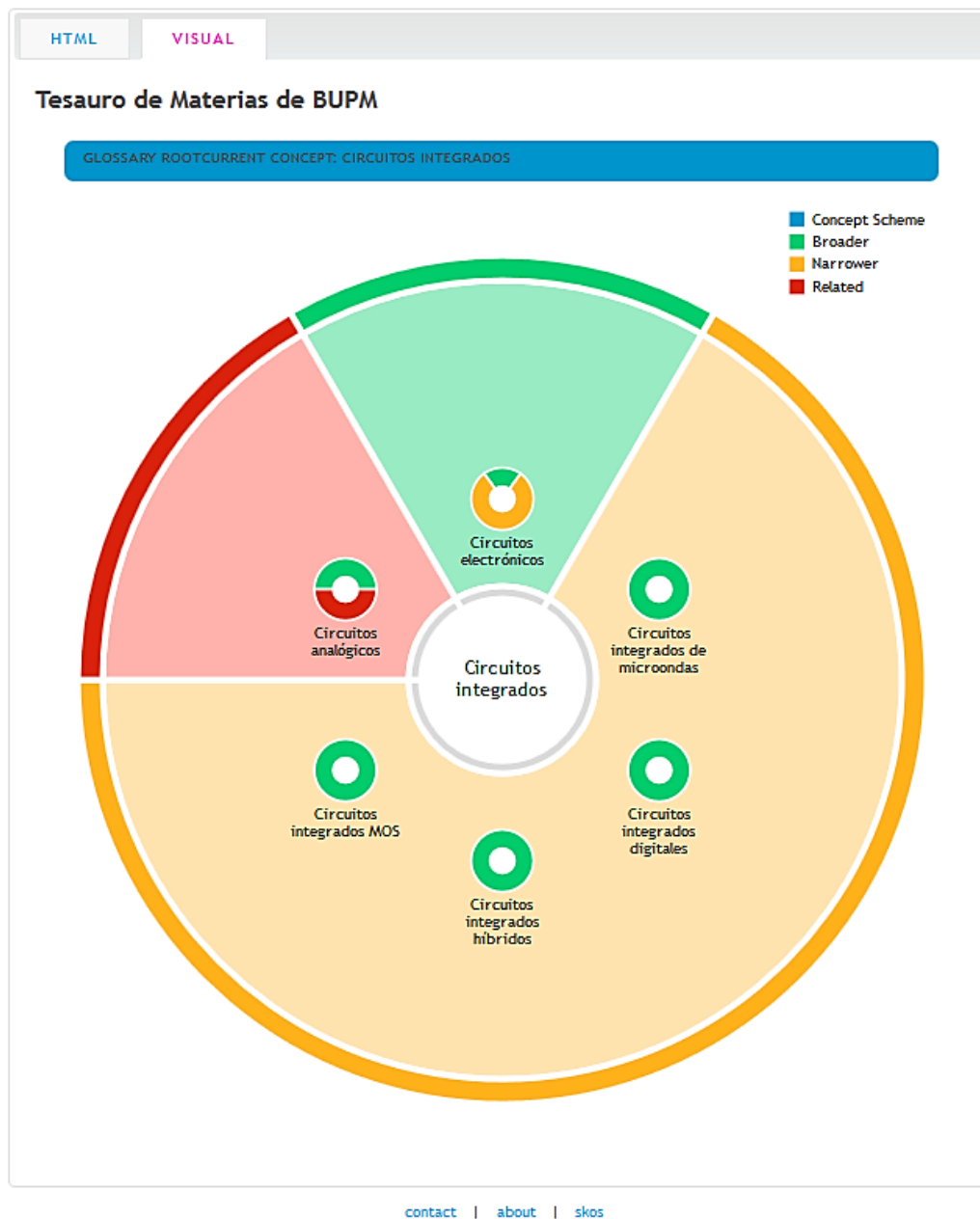


Figura 26 Detalle de visualización de conceptos en Tesouro BUPM. Fuente: (Ávila, 2014c).

6.1.2 PUBLICACIÓN EN THE DATAHUB

La publicación de datos en forma de “*data dumps*” es otra de las posibilidades con la que contamos para aumentar la visibilidad y usabilidad del dataset. The DataHub es la plataforma de referencia para ello, ofreciéndonos un entorno compatible e interoperable que sigue las pautas de *LOD Cloud* y de las recomendaciones W3C sobre Linked Data y bibliotecas.

The DataHub se basa en el software de publicación y gestión de datos CKAN, quizás el sistema más utilizado para publicar datos propios. La publicación es libre de coste y no requiere grandes conocimientos técnicos. Desde el punto de vista de la visibilidad, la presencia de los datos en la plataforma aumenta exponencialmente las posibilidades de reutilización, para ello se cuenta con herramientas eficaces de búsqueda que rastrean contenidos apoyándose en varios criterios, uno de los más relevantes es la búsqueda por metadatos y facetas generales. Especialmente importante en la presencia de metadatos de calidad, imprescindibles para los rastreos automáticos que realizan aplicaciones de búsqueda semántica en la plataforma (Cyganiak & et al., 2011; Open Knowledge Foundation, 2014a; Open Knowledge Foundation, 2014c).

Para aumentar la visibilidad del proyecto, que aquí se presenta se ha abierto un espacio propio en la plataforma para almacenar la totalidad de ficheros generados en el proyecto. El conjunto datos se ha denominado “Tesauro materias BUPM” incluido en la organización “Vocabularios Linked Data para bibliotecas”.

El proceso de alta no es muy complejo:

1. Se requiere en primer lugar efectuar un registro y elegir o crear una “organización” (para CKAN una organización es una entidad de gestión de datasets, muy parecido en posibilidades a los catálogos de datos), para la cual crearemos un título, una descripción y un vínculo. La organización creada o a la que nos hemos adherido, tendrá que tener ajustados unos privilegios de acceso, evidentemente en caso de adhesión a una organización esas reglas nos vendrán dadas. Desde hace unos meses la posibilidad de creación de “organizaciones” no es automática en The DataHub, se requiere contacto con los administradores de la plataforma, aportar un subdominio para la organización, y credenciales de acceso.
2. Posteriormente se ha de registrar el dataset cumplimentando datos como el título, la licencia, el link de descarga, etc.
3. La publicación del dataset puede incorporar cualquier elemento que creamos conveniente para favorecer la reutilización, no existiendo, en principio limitaciones de formatos, aunque se recomiendan modelos semánticos de representación de datos.

4. Cualquier información de la cuenta de usuario, de los datasets o los perfiles de las organizaciones son editables.

A partir de la publicación, el dataset está disponible para su reutilización por cualquiera que utilice The DataHub, bajo el único condicionante de la licencia de uso. Conviene por ello que tanto los ficheros de metadatos descriptivos, como las licencias no estén embebidos en los ficheros del dataset, pues esto impediría su consulta inmediata e incluso podrían pasar por no existentes. La plataforma ofrece una variedad de opciones de licencia que no siempre coinciden con la concreta elegida. Para una expresión más específica se puede referenciar la licencia mediante un URL. En el proyecto que aquí se presenta se publica el dataset de materias, los links a la interfaz de visualización del tesoro de materias, los ficheros de metadatos VoID y DCAT, el fichero de la licencia, el link al servicio SPARQL Endpoint y así como el archivo de empaquetado para preservación (Open Knowledge Foundation, 2014a; Open Knowledge Foundation, 2014c).

6.1.3 ESTRUCTURA DE FICHEROS Y RECURSOS QUE INTEGRAN LA PUBLICACIÓN DEL PROYECTO

En la tabla que sigue se listan todos los ficheros y recursos que integran este proyecto y que están alojados en el repositorio de datos CKAN. También se incluye el enlace a la interfaz web del tesoro de materias.

Tabla 8 Referencia a los ficheros que integran el Tesoro de materias BUPM

| |
|---|
| Página principal del Tesoro de materias BUPM (Acceso) |
| Fichero TRIG Tesoro de materias BUPM (Acceso) |
| Metadatos VoID (Acceso) |
| Metadatos Provenance (Acceso) |
| Metadatos DCAT (Acceso) |
| Declaración de licencia CC-BY-4.0 (Acceso) |

SPARQL Endpoint ([Acceso](#))

Declaración de licencia CC-BY-4.0 ([Acceso](#))

Fichero de empaquetado de recursos para la preservación ([Acceso](#)) ([Visualización](#))

6.2 VOCABULARIOS DE METADATOS DESCRIPTIVOS PARA LINKED DATASETS

6.2.1 VOCABULARY OF INTERLINKED DATASETS (VoID)

VoID es uno de los lenguajes RDF específicamente diseñados para la asignación de metadatos descriptivos a datasets, ya desde un punto de vista local ya desde la posible interconexión de vocabularios mediante mapeos. Este vocabulario de metadatos mejora la identificación del dataset, permitiendo al usuario conocer mejor sus características y optimizando la selección del grupo de datos a utilizar.

Su estructura de descripción tiene varios niveles: en primer lugar se establece un formato de descripción basado en vocabularios de metadatos generalistas como Dublin Core o FOAF. Estos metadatos identifican al conjunto de datos, suministran información de acceso y del modelo de estructura. En segundo lugar se describe la vinculación entre uno o varios vocabularios, en nuestro caso con LCSH, RAMEAU, SWD y Nuovo Soggettario. Las ventajas son evidentes: descubrimiento más fácil de los datos en las búsquedas por la declaración de la forma de acceso, perfecta identificación de licencias de uso, atribución correcta de la propiedad intelectual de los datos y facilidad para el reutilizador a la hora del posible alineamiento del vocabulario (Keith, Cyganiak, Hausenblas, & Zhao, 2011).

Las dos principales clases del vocabulario son `void:Dataset` y `void:Linkset`. VoID define “*dataset*” como un conjunto de tripletas RDF publicadas por un solo proveedor. El conjunto de datos es una entidad que identifica el conjunto de declaraciones RDF, permitiendo focalizar en ella el resto de declaraciones descriptivas del conjunto de datos. La clase `void:Linkset` sirve para establecer vínculos entre conjuntos de datos a través de triples RDF, es decir, relaciones de mapeo.

VoID puede proveer datos generales descriptivos utilizando sus propias propiedades (especificaciones técnicas, `void:features`) o vocabularios como DCMI terms para suministrar la información como el título, la descripción del dataset, los datos de contacto del que publica, el

tema o materia, la licencia asignada; o FOAF, para especificar la web base del proyecto o propiedades para definir y asignar la licencia adecuada. Estos metadatos ayudan al usuario a decidir si un conjunto de datos determinado es apropiado para sus propósitos.

Los metadatos de acceso permiten encontrar los triples contenidos vía resolución de IRIs: directamente mediante declaración de SPARQL endpoint, a través de la declaración de búsqueda de texto libre en el conjunto de datos `void:openSearchDescription` o de un recurso de entrada principal a los datos a través de `void:rootResource`, propiedad de especial importancia en el caso de vocabularios jerárquicos (Cyganiak et al., 2010).

VOID también proporciona metadatos de descripción de la estructura interna y del esquema del dataset, como vocabularios incluidos, ejemplos de recursos contenidos, patrones de construcción de IRIs, descripciones de partes del conjunto de datos.

La descripción de vinculación entre conjuntos de datos se establece mediante `void:Linkset` y supone que existe una colección de tripletas cuyos sujetos están en el dataset origen y los objetos en el conjunto de datos vinculado. La propiedad opcional que establece el tipo de unión se expresa mediante `owl:sameAs`. La propiedad `void:target` sirve para identificar al conjunto de datos objetivo (Keith et al., 2011). En el ejemplo siguiente se expone la descripción VOID del prototipo de tesauro de materias.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/> a void:Dataset;
<http://datahub.io/es/dataset/tesouro-materias-bupm/> a void:Datarump;

#Descripción#
dcterms:title "Encabezamiento de Materia Biblioteca UPM";
dcterms:contributor :Biblioteca Universitaria Campus Sur;
dcterms:modified "2014-05-25"^^xsd:date;
foaf:homepage <http://academia6.poolparty.biz/Tesouro-Materias-BUPM/>;
dcterms:publisher "Rafael Ávila";
foaf:mbox <mailto:rafael.avila@xxupm.es>;
dcterms:subject dbpedia:Subject_headings;
dcterms:license < href="http://creativecommons.org/licenses/by/4.0/>;
void:feature <http://www.w3.org/ns/formats/Turtle>;
void:sparqlEndpoint <http://academia6.poolparty.biz/PoolParty/sparql/Tesouro-
Materias-BUPM >;
void:exampleResource <http://upm.es/biblioteca/kos/sh/Lenguajes de
programación>,
<http://academia6.poolparty.biz/Tesouro-Materias-BUPM/273>;
void:vocabulary <http://www.w3.org/2004/02/skos/core/> ,
<http://www.loc.gov/mads/rdf/v1#> ,
<http://www.w3.org/2008/05/skos-xl#> ;
```

```

<http://upm.es/biblioteca/kos/sh/bibupm-LCSH> a void:Linkset;

<http://upm.es/biblioteca/kos/sh/bibupm-LCSH> a void:Linkset;
<http://upm.es/biblioteca/kos/sh/bibupm-RAMEAU> a void:Linkset;
<http://upm.es/biblioteca/kos/sh/bibupm-SWD> a void:Linkset;
<http://upm.es/biblioteca/kos/sh/bibupm-Nuovo-Soggiretario> a void:Linkset;
void:target <http://upm.es/biblioteca/kos/sh/UPMSubjectHeadings>;
void:target < http://data.bnf.fr/liste-rameau>;
void:target < http://thes.bncf.firenze.sbn.it/ricerca.php>;
void:target <http://id.loc.gov/authorities/subjects>;
void:target < https://portal.dnb.de/opac.htm>;
void:rootResource <http://academia6.poolparty.biz/Tesauro-Materias-BUPM/> .

#URL Declaración VoID"
https://ckannet-storage.commondatastorage.googleapis.com/2014-07-
28T10:28:42.182Z/void-bupm-sh.ttl
#Encabezamientos de Materia Biblioteca UPM .

```

Existen otras posibilidades de declaración de metadatos. Podemos utilizar para ello RDF complementado con algún esquema de elementos de metadatos como DC, FOAF, etc. Existe una evolución no normativa de VoID en combinación con FRBR para la descripción de datos desarrollada por Bernard Vatant denominada VOAF “*Vocabulary of a Friend*” cuyo objetivo además de la propia descripción es el de generar redes de vocabularios relacionados. VOAF se considera una subclase de VoID y de FRBR, complementando a ambos y añadiendo información especialmente orientada a la agrupación semántica de vocabularios (Vatant, 2013).

6.2.2 DATA CATALOG VOCABULARY (DCAT)

Data Catalog Vocabulary (DCAT) es una recomendación de la W3C (2014) para la descripción de catálogos de datos y datasets con el objetivo de hacerlos más interoperables e identificables. Las descripciones de datos con DCAT mejoran la disponibilidad y visibilidad de los datos, facilitan el proceso a las aplicaciones y asignan valores imprescindibles para la preservación. DCAT tiene como función principal describir los catálogos, los datasets que los integran y aportar información sobre los publicadores de datos, especialmente en el contexto de los catálogos Open Data en las publicaciones del ámbito eGovernment. (Maali & Erickson, 2014).

DCAT va a ser utilizado aquí a los únicos efectos de descripción del conjunto de datos, no existe un catálogo al que podamos adscribirlo. DCAT es el estándar de descripción que prescribe la Norma Técnica de Interoperabilidad (NTI 2013), lo cual obliga a clasificar los datasets según una taxonomía temática determinada según los catálogos de sectores primarios donde se encuadran los diferentes tipos de actividad del ente que publica. En nuestro caso el vocabulario se adscribe a la categoría:

`http://datos.gob.es/kos/sector-publico/sector/educacion`

Se describen a continuación, brevemente, los principales elementos del vocabulario. La clase `dcat:Catalog` representa al catálogo y asigna metadatos generales sobre el mismo y los datasets que lo componen. Sus principales propiedades son `dcat:dataset` que identifica al dataset como parte del catálogo, `dcat:record`, que identifica un registro como parte del catálogo y `dcat:themeTaxonomy`, que especifica las áreas temáticas que se utilizan para clasificarlo. El resto de propiedades son elementos DC terms como: `language`, `description`, `title`, `issued`, `rights`, `license`, `spatial`, que estructuran valores de descripción del catálogo.

La clase `dcat:Dataset` representa al conjunto de datos en referencia a su publicación por un editor determinado y descargable en varios formatos. Sus principales propiedades son: `dcat:theme`, que identifica el principal tópico del dataset, `dcat:keyword`, que asigna una etiqueta identificativa al dataset, `dcat:distribution` que enlaza los datasets con sus distribuciones, `dcat:landingPage`, la página web que permite el acceso al dataset o a sus distribuciones. Utiliza las mismas propiedades DC terms que la clase `dcat:Catalog`,

La clase `dcat:Distribution` define un determinado formato de representación del dataset. Sus principales propiedades son: `dcat:accessURL` que indica el recurso (web, SPARQL Endpoint, etc) que permite el acceso a la distribución del dataset, `dcat:downloadURL`, indica el fichero que contiene la distribución, `dcat:byteSize`, el tamaño en bytes de la distribución, `dcat:mediaType`, el tipo de medio definido ((Freed, 2014)). Utiliza las habituales propiedades descriptivas de DC terms.

La clase `dcat:CatalogRecord`, se utiliza para distinguir la información sobre el dataset propiamente dicha, respecto de la información que se genera cuando el conjunto de datos se añade a un catálogo determinado. Sus principales propiedades provienen de DC terms: `title`, `description`, `issued` o `modified`.

Finalmente, DCAT, utiliza algunas clases principales del espacio de nombres de SKOS y FOAF (en este caso fundamentalmente para la descripción de personas y organizaciones) (Maali & Erickson, 2014).

Como se dijo anteriormente, se entiende que este proyecto está incluido en el ámbito normativo del RD 1495/2011 de desarrollo de la Ley 37/2007 sobre Reutilización de la Información del Sector Público. En este sentido la Norma Técnica de Interoperabilidad de reutilización de recursos de información (NTI), define las pautas para la asignación de metadatos descriptivos que deben acompañar a la información reutilizable y la obligatoriedad de la inclusión de esa información tanto en catálogos de datos como en datasets. La NTI indica una serie de vocabularios preceptivos para la descripción: DCAT, DC , SKOS, XML Schema, W3C Time Ontology y FOAF. En la siguiente figura se define el modelo de datos DCAT. (Ministerio de Hacienda y Administraciones Públicas, 2013):

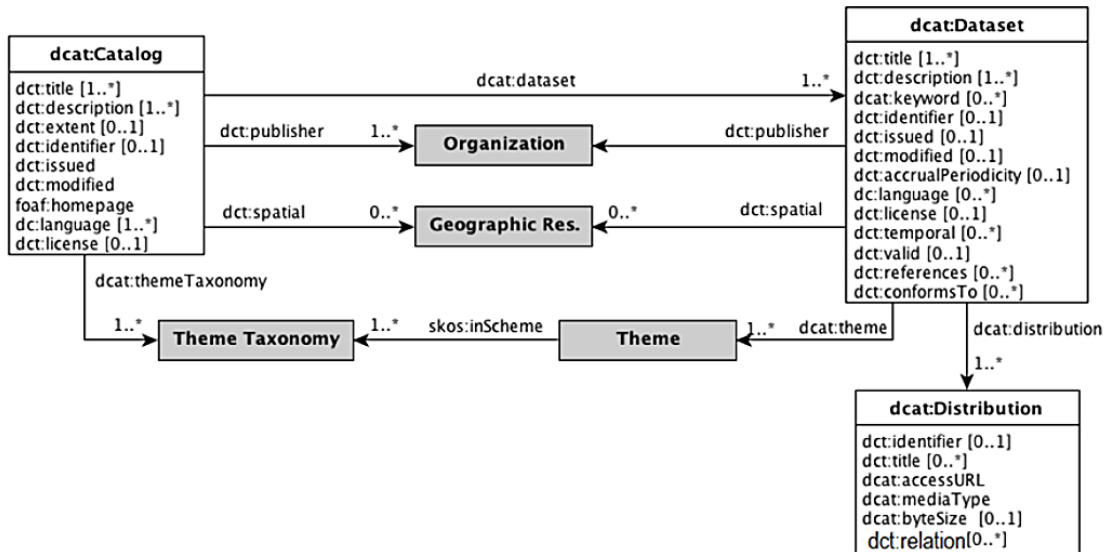


Figura 27 Modelo de datos Data Catalog Vocabulary (DCAT). Fuente: (Ministerio de Hacienda y Administraciones Públicas, 2013)

El vocabulario de materias de BUPM se describirá como una instancia de `dcat:Dataset`:

```

@prefix dct: <http://purl.org/dc/terms/>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix time: <http://www.w3.org/2006/time#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

#Todas las IRIs y URLs referidas a UPM son ficticias#
<http://academia6.poolparty.biz/Tesauro-Materias-BUPM/> a dcat:Dataset;
dct:title "Catálogo de vocabularios de la Universidad Politécnica de Madrid"@es;
dct:description "Catálogo de vocabularios en Linked Data de la Biblioteca de la Universidad Politécnica de Madrid"@es;
#valor escogido de la taxonomía general del sector público (NTI)#
dcat:theme <http://datos.gob.es/kos/sector-publico/sector/educación>;
dcat:keyword "vocabularios","Linked Data","biblioteca";
dct:accrualPeriodicity
[
  a dct:Frequency;
  rdf:value "P7D"^^xsd:timePeriod ;
  rdfs:label "Cada semana".
];
dct:publisher <http://www.upm.es/biblioteca/>;
dct:identifier "http://www.upm.es/catalogo/XXA567";
dct:issued "2014-07-10"^^xsd:date;
dct:modified "2014-07-15"^^xsd:date;
dc:language "es";
  
```

```

dct:license <"http://creativecommons.org/licenses/by-nc/4.0/"> ; dct:spatial
< http://www.geonames.org/8030677/>;
dct:temporal
[
  a dct:PeriodOfTime, time:Interval;
  time:hasBeginning
  [
    a time:Instant;
    time:inXSDDateTime "2014-07-10"^^xsd:date.
  ];
  time:hasEnd
  [
    a time:Instant;
    time:inXSDDateTime "2014-09-15"^^xsd:date.
  ].
];
dcat:distribution <http://academia6.poolparty.biz/Tesauro-Materias-BUPM/> ;
dcat:distribution <https://ckannet-storage.commondatastorage.googleapis.com/2014-07-
28T18:49:19.579Z/datapackage.json> .

```

6.2.3 PUBLICACIÓN DE DATOS DE LAS ADMINISTRACIONES PÚBLICAS. AJUSTE CON LA NORMA TÉCNICA DE INTEROPERABILIDAD

La Norma Técnica de Interoperabilidad en su apartado VIII (Ministerio de Hacienda y Administraciones Públicas, 2013) va a regular la publicación de la información de un modo reutilizable. Establece un marco presidido por el “principio de accesibilidad a la información y a los servicios” por medios electrónicos, y lo hace garantizando la accesibilidad universal a los recursos expuestos sea cual sea los condicionantes previos de quien va a utilizar los datos. Regula la necesidad de ofrecer los productos reutilizables en sites propios, lo que facilita su localización y uso.

Como hemos señalado anteriormente, la publicación de información pública de forma estructurada debe ir acompañada de una descripción válida y suficiente de su contenido, y debe hacerlo mediante sistemas prescritos en la propia norma (conjunto de vocabularios de descripción de recursos subsumidos en el estándar DCAT). La garantía de reutilización incluye la divulgación de documentación de ayuda a los agentes reutilizadores: formatos utilizados, sistemas de acceso, configuración de navegadores, instrucciones de consulta, etc.

El apartado IX de la NTI define las normas de publicación de información reutilizable. Se debe ofrecer una interfaz de publicación y de consulta de datos que permita cumplir los objetivos de la reutilización. Para ello se utilizarán estándares de estructuración de metadatos (la propia guía ofrece ejemplos en RDF de descripciones de metadatos de los catálogos, los datases y las distribuciones de los mismos). La publicación se debe efectuar mediante documentos HTML que contengan las descripciones semánticas de los datos (Ministerio de Hacienda y Administraciones Públicas, 2013). En la figura inferior se puede observar un modelo básico de interacción: editores, datos y utilizadores, propuesto por la NTI:

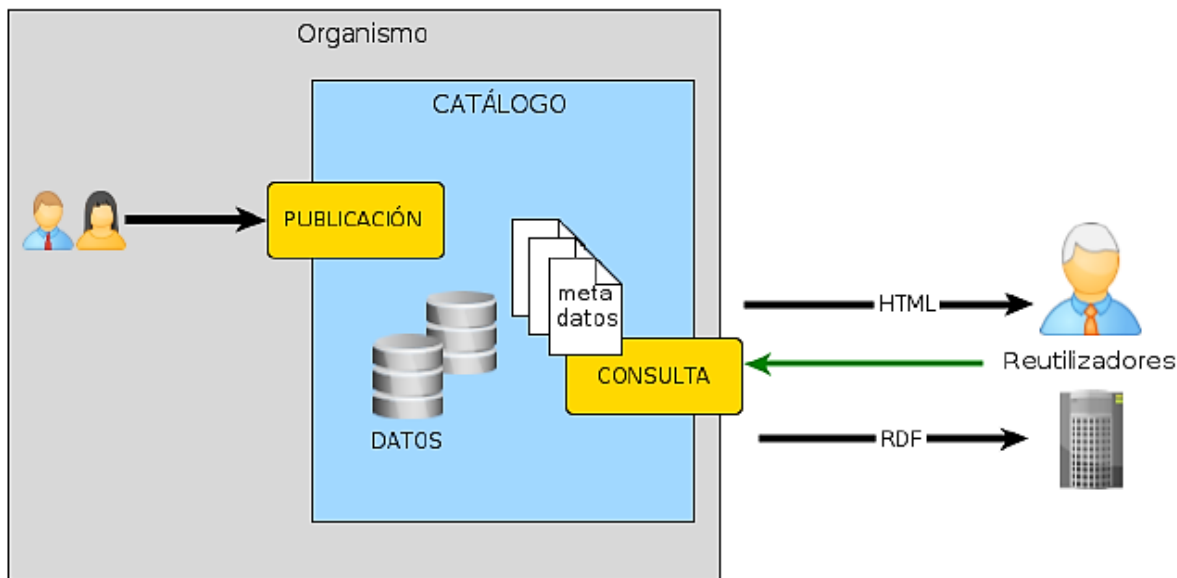


Figura 28 Proceso de interacción de los agentes participantes en la cadena de datos de las administraciones públicas. Fuente: (Ministerio de Hacienda y Administraciones Públicas, 2013).

6.2.4 VOCABULARIO PARA LA DESCRIPCIÓN DE LA PROCEDENCIA. PROVENANCE

Uno de los principales objetivos que persigue la descripción de los datos mediante “*Provenance*” es el de expresar el origen de los datos, su evolución y quién y cómo se construyeron la estructuras de datos. Como principio, se trata de ofrecer datos de calidad y fiables, con la garantía de tener en nuestra mano información relevante para la evaluación de los datos e investigar hacia el pasado fuentes de información de nuestro interés. Estos datos relevantes para definir correctamente la procedencia provienen de los proveedores de recursos LOD y de los metadatos técnicos que suministran algunas aplicaciones (Gil & Miles, 2013).

La Web de datos contiene multitud de elementos interconectados cuyas funciones pueden desplegarse en un amplio ámbito: describiendo recursos, estructurando valores, ordenando y filtrando datos. La propia naturaleza de Linked Data puede originar que idénticos recursos o set de datos estén apuntados por diferentes declaraciones. Distinguir entre esa variedad de recursos es el principal objetivo de “*Provenance*”, aportando métodos para detectar la fiabilidad, pero sobre todo, información sobre qué recurso es el más actualizado, o cuya fuente es de más reconocido prestigio y por ello confiable. (Dunsire & Willer, 2013; Hartig & Zhao, 2010).

El vocabulario del W3C “*Provenance*” permite describir el flujo temporal de los datos, sus transformaciones y los agentes que han participado en cada parte. Esta información es de crucial importancia y debe asociarse al dataset al mismo nivel que otras declaraciones como licencias o datos de preservación. Sin embargo no está generalizada su inclusión como complemento a las

publicaciones de datos, cuestión que debe hacer reflexionar a la comunidad semántica que debe promover procesos de publicación normativos completos. Algunos autores defienden que “*Provenance*” debe asimilarse en categoría a los principios Linked Data, argumentando para ello que la publicación masiva de datos vinculados pierde parte de su valor y posibilidad de reutilización, si carece de ese tipo de información (Gil & Miles, 2013).

El vocabulario “*Provenance*” recoge información sobre los participantes en la generación de la información: sobre entidades (recursos, conceptos o cosas en general), actividades (procesos de las cosas u objetos descritos), agentes (se refiere a quién interviene en la creación o uso de entidades o actividades), roles (indican la función de una entidad en una actividad). En la figura siguiente se muestra el modelo de datos básico de “*Provenance*” (Gil & Miles, 2013).

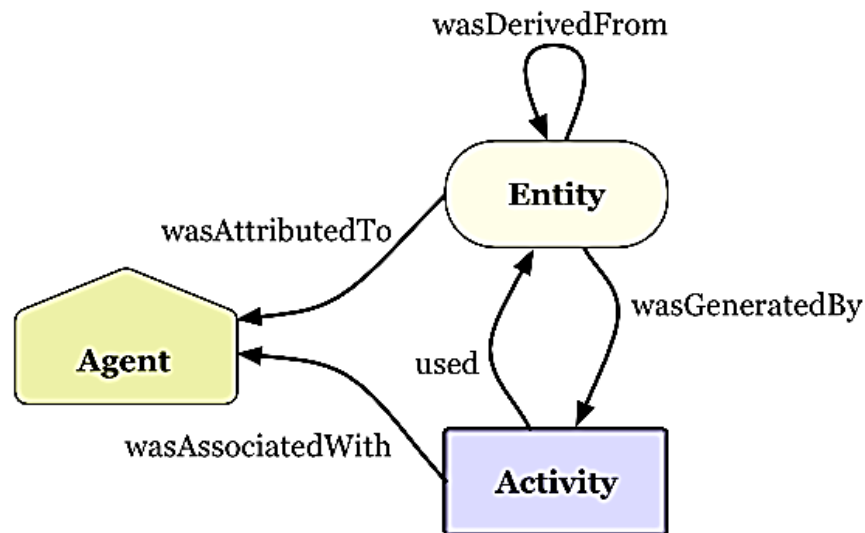


Figura 29 Modelo de datos Provenance. Fuente:(Gil & Miles, 2013)

El vocabulario “*Provenance*” puede utilizar otros “*namespaces*” para la asignación de características y valores a las entidades, actividades y agentes.

Se describe a continuación un proceso simple de modelado del tesoro de materias mediante su información de procedencia, este fichero se incorporará en el empaquetado del proyecto. Para la expresión de sujetos y objetos de las tripletas se utilizan “*namespaces*” no reales (<http://www.upm.es/prov/sh>; <http://www.upm.es/prov/umporg>), cuyos prefijos son. “*provsh:*” para nombres de entidades y actividades y “*umporg:*” para los agentes. Se definen entidades para cada paso en la evolución del vocabulario de materias (*provsh:prodataset1.....*), se identifican las sucesivas etapas de desarrollo del tesoro, la

relación entre entidades y actividades (una entidad usada por una actividad genera una nueva entidad y así sucesivamente), se describe el agente publicador, las dependencias entre actividades y entidades, el agente planificador y supervisor y el producto final del proceso.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix sh: <http://www.upm.es/biblioteca/kos/sh#> .
@prefix upmorg: <http://www.upm.es/biblioteca/org#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
```

#Se utilizan espacios de nombres ficticios para identificar los productos del proyecto en evolución y para identificar a la organización#

#Entidades#

#Descripción de entidades según la evolución del proyecto#

```
provsh:prodataset1 rdf:type prov:Entity ;
dcterms:title "Encabezamientos de materia Biblioteca UPM 1.0" .
provsh:prodataset2 rdf:type prov:Entity .
dcterms:title "Encabezamientos de materia Biblioteca UPM 2.0" .
provsh:prodataset3 rdf:type prov:Entity . dcterms:title "Encabezamientos de
materia Biblioteca UPM 3.0" . provsh:publishdataset rdf:type prov:Entity .
dcterms:title "Encabezamientos de materia Biblioteca UPM Versión final" .
```

Actividades#

#Descripción de las actividades por etapas en la generación del tesauro de materias#

```
provsh:classification rdf:type prov:Activity .
provsh:generation rdf:type prov:Activity .
provsh:publication rdf:type prov:Activity .
```

#Relaciones de actividades y los productos intermedios de generación del tesauro de materias#

#Generación de nuevas entidades #

```
sh:clasification prov:used sh:prodataset1 . sh:prodataset2
prov:wasGeneratedBy sh:classification . sh:generation prov:used
sh:prodataset2 . sh:prodataset3 prov:wasGeneratedBy sh:generation .
sh:publication prov:used sh:prodataset3 . sh:publishdataset
prov:wasGeneratedBy sh:dataset .
```

Agentes

```
sh:classification prov:wasAssociatedWith upmorg:SBU-UPM .
Upmorg:Servicio de Biblioteca Universitaria UPM rdf:type prov:Agent;
rdf:type prov:Organization;
foaf:name "Servicio de Biblioteca Universitaria UPM.
```

```

sh:generation prov:wasAssociatedWith upmorg:Avila-R .
sh:publication prov:wasAssociatedWith upmorg:Avila-R .
upmorg:Avila-R rdf:type prov:Agent ;
rdf:type prov:Person ;
foaf:name "Rafael Ávila" .

# Revisiones y subordinaciones de entidades

sh:prodataset2 rdf:type prov:Entity ;
prov:wasDerivedOf sh:prodataset1 .
sh:prodataset3 rdf:type prov:Entity ;
prov:wasDerivedOf sh:prodataset2 .
sh:publishdataset rdf:type prov:Entity ;
prov:wasRevisionOf sh:prodataset3 .

# Planificación
sh:revision rdf:type prov:Activity .
upmorg:Ortiz-V rdf:type prov:Agent;
rdf:type prov:Person ;
foaf:name "Virginia Ortiz Repiso" .
sh:instructions rdf:type prov:Plan .
sh:revision prov:qualifiedAssociation
[ a prov:Association ;
  prov:agent upmorg:Ortiz-V ;
  prov:hadPlan sh:instructions
] .
sh:publishdataset prov:wasGeneratedBy sh:prodataset3 .

# Fecha de publicación#
sh:publishdataset prov:generatedAtTime "2014-05-20T10:30:00"^^xsd:dateTime .

```

7 CONSUMO DE DATOS Y PRESERVACIÓN LINKED DATA

Llegamos al punto final donde se concretan todos los trabajos anteriores. Las consideraciones de consumo pueden abordarse desde diferentes puntos de vista, el que aquí vamos a utilizar es aquel más cercano a las expectativas bibliotecarias.

El análisis de posibles usuarios del producto tiene en cuenta la reutilización en cualquier contexto, pero las posibilidades de uso aumentan si se enfocan los usos profesionales que el vocabulario de materias pueda tener. Se pretende pues ofrecer un producto básico en forma de vocabulario multilingüe, cuyo consumo satisfaga los niveles de calidad de cualquier producto bibliotecario y las expectativas previstas para ordenar el conocimiento en cualquier organización. La estrategia seguida es seguir las normas más actuales tanto para la reestructuración del vocabulario como para su mapeo, ofreciendo interfaces de consulta y visualización suficientes en la medida de las posibilidades de este proyecto.

Por otro lado se ha seguido en lo posible, las mejores prácticas de vocabularios de prestigio en el ámbito de los encabezamientos de materia, intentando implementar las lecciones aprendidas por las organizaciones que los construyeron. Para ello no se ha considerado necesario realizar un análisis previo y en profundidad de los vocabularios disponibles, a través por ejemplo de búsquedas en Síndice o el propio The DataHub; el prestigio algunas de las principales bibliotecas del mundo y la literatura de referencia han sido suficientes para escoger los mejores y más completos (y diversos) vocabularios.

Para el fomento de consumo se han elegido las licencias más adecuadas, a nuestro juicio, ofreciendo todas las facilidades para la reutilización. También se suministra una interfaz de consulta SPARQL, además de la propia interfaz interactiva del tesoro y como complemento se sirven los ficheros completos en la plataforma The DataHub.

Se establece un compromiso sobre el posible consumo futuro del proyecto, intentando mantener la integridad del vocabulario, manteniendo en buen estado el sistema de mapeos y previendo también la conservación de los propios datos. A nuestro juicio, se debería implementar un sistema de vigilancia tecnológica que monitorizara las rápidas evoluciones tecnológicas, cuestión que no se ha definido en este proyecto por carácter de prototipo.

Se va a tratar a continuación como epígrafes fundamentales del consumo las posibilidades de búsqueda mediante el lenguaje SPARQL y el crucial tema de la asignación de licencias. En otras partes del trabajo se han definido otros productos integrantes de la estrategia de consumo: desde el análisis de recuperación de la información, las interfaces y plataformas de publicación y el sistema de metadatos descriptivos que informan de las posibilidades de uso y los contenidos del vocabulario.

7.1 SPARQL PROTOCOL AND RDF QUERY LANGUAGE

SPARQL (SPARQL Protocol And RDF Query Language) es un lenguaje de consulta para la obtención de información a partir de conjuntos de datos RDF; también es un protocolo que permite especificar comandos SPARQL, permitiendo la transmisión de consultas y respuestas entre un cliente y un motor de SPARQL bajo HTTP, lo que se efectúa, habitualmente, a través de una aplicación de servidor denominada SPARQL Endpoint. Como complemento existe una extensión del lenguaje denominada SPARQL Update la cual permite una gestión avanzada de los datasets, pues puede utilizarse, en combinación con técnicas de búsqueda y filtrado, para añadir, revisar o eliminar partes de las tripletas recuperadas (Gearon, Passant, & Polleres, 2013; Harris & Seaborne, 2013).

La sintaxis básica del lenguaje de consulta SPARQL se estructura en varias áreas definidas:

1. Declaración de los prefijos e IRIs utilizados en el dataset.
2. Declaración del conjunto de datos en cuyo ámbito se ejecutará la búsqueda.
3. Formatos de consulta que utilizan patrones de coincidencia para expresar los resultados.

Existen cuatro tipos:

- a. SELECT: devuelve conjuntos de variables según el patrón de búsqueda.
 - b. CONSTRUCT: devuelve un grafo RDF que sustituye a las variables.
 - c. ASK: devuelve el valor booleano que indica coincidencia o lo contrario.
 - d. DESCRIBE: devuelve un grafo RDF que describe los recursos.
4. Patrón de búsqueda (WHERE), que no es sino un tipo especial de declaración RDF con variables.
 5. Variables precedidas por el símbolo "?", que pueden contener cualquier nodo, recurso o literal del dataset consultado.
 6. Modificadores variables que se utilizan para filtrar, ordenar o realizar operaciones aritméticas, etc (Euclid, 2013; Feigenbaum & Prud'hommeaux, 2014; Harris & Seaborne, 2013).

Ejemplo de dataset de referencia. Se trata de las declaraciones implicadas en mapeos entre vocabularios. Se utilizan la serialización de las materias de la BUPM y su vinculación con el concepto equivalente en el vocabulario objetivo (LCSH, RAMEAU, SWD).

```
<http://upm.es/biblioteca/kos/sh/0546898> a skos:Concept ;  
skos:prefLabel "Programas de computador"@es .  
<http://upm.es/biblioteca/kos/sh/0546898> skos:closeMatch  
<http://data.bnf.fr/ark:/12148/cb133183707> .  
<http://data.bnf.fr/ark:/12148/cb133183707> a skos:Concept ;
```



```
skos:prefLabel "Logiciels"@fr .
```

```
<http://upm.es/biblioteca/kos/sh/0546898> skos:closeMatch  
<http://id.loc.gov/authorities/subjects/sh85029524> .  
<http://id.loc.gov/authorities/subjects/sh85029524> a skos:Concept;  
skos:prefLabel "Computer programs"@en .
```

```
<http://upm.es/biblioteca/kos/sh/0546898> skos:closeMatch  
<http://d-nb.info/gnd/4047394-6> .  
<http://d-nb.info/gnd/4047394-6> a skos:Concept;  
skos:prefLabel "Programm"@de .
```

#Ejemplo de consultas SPARQL:#

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>  
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>  
SELECT ?idioma  
FROM <http://upm.es/biblioteca/kos/sh#>  
FROM <http://data.bnf.fr/ark:/ >  
FROM <http://d-nb.info/gnd/>  
FROM <http://id.loc.gov/authorities/subjects/>
```


```
WHERE {  
  ?x skos:prefLabel ?y@en .  
}
```

La consulta a los dataset recuperaría las materias cuyo idioma fuera el inglés.

<http://id.loc.gov/authorities/subjects/sh85029524>

[http://id.loc.gov/authorities/subjects/Computer programs@en](http://id.loc.gov/authorities/subjects/Computer%20programs@en)

La plataforma de gestión y publicación de tesauros PoolParty, utilizada para este proyecto ofrece un servicio de consulta mediante un SPARQL EndPoint. En las ilustraciones siguientes se presenta una consulta en la que se pretende que el sistema recupere todas los IRIs cuyas materias contengan el concepto “circuito” como etiqueta preferente.



Tesouro de Materias de BUPM

[Wiki](#)
[SPARQL](#)
[Logout](#)

SPARQL Endpoint

```

PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?Concept ?prefLabel
WHERE
{
  ?Concept ?x skos:Concept .
  { ?Concept skos:prefLabel ?prefLabel . FILTER (regex(str(?prefLabel), '^circuito.*', 'i')) }
} ORDER BY ?prefLabel LIMIT 50 OFFSET 0

```

Query valid!

Format: HTML Table Run Query

Add Namespace

☐ SKOS
 ☐ DC
 ☐ DCTERMS
 ☐ OWL
 ☐ RDF
 ☐ RDFS
 ☐ SWC

Test

[Sample Query 1](#)
 Returns all URIs and preferred labels of concepts that start with the letter "A" and sorts them alphabetically. A maximum of 50 concepts are displayed.

[Sample Query 2](#)
 Returns 50 Triples of any kind.

[Sample Query 3](#)
 Returns preferred and alternative label and the definition of a maximum of 50 concepts where these values are defined.

[contact](#)
[about](#)
[skos](#)

Figura 30 Detalle de ecuación de búsqueda en cliente SPARQL Endpoint. Fuente: (Ávila, 2014a).

| Concept | prefLabel |
|--|---|
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/328 | "Circuitos analógicos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/330 | "Circuitos de impulsos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/320 | "Circuitos de transistores"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/319 | "Circuitos electrónicos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/316 | "Circuitos eléctricos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/318 | "Circuitos eléctricos lineales"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/317 | "Circuitos eléctricos no lineales"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/321 | "Circuitos impresos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/322 | "Circuitos integrados"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/326 | "Circuitos integrados MOS"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/323 | "Circuitos integrados de microondas"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/324 | "Circuitos integrados digitales"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/325 | "Circuitos integrados híbridos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/331 | "Circuitos lógicos"@es |
| http://academia6.poolparty.biz/Tesouro-Materias-BUPM/332 | "Circuitos lógicos integrados"@es |

Figura 31 Resultados de la búsqueda de la figura 30. Fuente: elaboración propia

7.2 LICENCIAS PARA DATOS VINCULADOS

7.2.1 Aspectos generales para la asignación de licencias

Uno de los debates constantes en el contexto de la Web de datos consiste en la reflexión sobre el despegue definitivo de estas tecnologías. En este contexto se estudia el papel dinamizador que las organizaciones privadas pueden ofrecer al entorno Linked Data, uno cuyos efectos esperados es la mayor presencia de los temas legales. A veces el excesivo entusiasmo en la publicación impide reflexionar sobre las posibles consecuencias jurídicas y los posibles daños económicos que puede generar la utilización de datos sin la debida protección. La seguridad jurídica es la puerta de entrada de agentes dispuestos a disponer sus datos en la Web y expresar correctamente las licencias una necesidad estructural de la publicación de datos vinculados y de su consumo.

La primera estrella de la evaluación del sistema Linked Open Data exige la aplicación de una licencia abierta a los datos, la cual ofrezca información sobre sus posibilidades de utilización. La indicación exacta de qué se puede hacer o no con los datos, fomenta el uso y la reutilización. El usuario tiene la seguridad de que los datos y metadatos que necesita pueden ser usados (o no) bajo las directrices expuestas en la licencia de uso y además, si se pide atribución, la licencia puede ser un buen instrumento para promocionar nuestro trabajo.

Podemos diferenciar dos grandes modos de licenciar los datos vinculados: datos que poseen una licencia de uso con mayor o menor amplitud de utilización, y datos sin licencia y por ello se entiende cerrados al uso abierto y en cuya categoría se identifican los denominados Linked Closed Data o Linked Enterprise Data. (Rodriguez-Doncel, Gomez-Pérez, & Mihindukulasooriya, 2013)

La asignación de una licencia, sea cual sea su rango de utilización, requiere un previo análisis de los datos que se quieren publicar:

1. Existencia o no de datos de carácter confidencial.
2. Marco jurídico aplicable a los datos según la legislación de propiedad intelectual.
3. Derechos de protección “sui generis” de las bases de datos, pues realmente no se las considera objetos de la propiedad intelectual. Este punto puede ser de gran importancia, pues en no pocas ocasiones la expresión de datos vinculados no hace sino reflejar tanto el contenido como la estructura de una base de datos origen. Esto quizás no es tan obvio en un primer momento, pero en realidad los datos que aparecen publicados en la Web y que conforman redes de datos interconectados, suelen residir en sistemas de bases de datos, tradicionales o de última generación (recordemos el caso de las bases de datos No SQL y Big Data).
4. Protección de los datos privados incluidos en bases de datos, pues el marco legal ha de ser el mismo en su traslación a datos vinculados.

En definitiva la ley protege a estos efectos:

1. A los creadores de contenidos depositados en datasets.
2. A los creadores de esos conjuntos de datos, los cuales han efectuado operaciones de curación sobre los datos y los registros.
3. A las personas cuya información personal está incorporada a los datos.
4. A terceros cuyos derechos se vean conculcados si los datos fueran conocidos.

(Parlamento Europeo & Consejo de la Unión Europea, 1995; Parlamento Europeo & Consejo de la Unión Europea, 1996; Rodríguez-Doncel et al., 2013)

Linked Data contiene diferentes estructuras que deben ser analizadas desde puntos de vista específicos, verificando su nivel de protección. Dado que los datos individuales no están protegidos, las tripletas únicamente tendrán protección en tanto en cuanto contengan datos protegibles expresamente, como por ejemplo que sean confidenciales. El dataset tiene protección en tanto en cuanto se asimila a una base de datos (aquí se debe tener en cuenta lo dicho sobre el derecho “sui generis” de las bases de datos, no existiendo consenso internacional sobre ello). La agregación de contenidos a los datasets supondrá siempre el respeto a la licencia de los datos agregados al uso. Los mapeos RDF se asimilan a una creación protegible y deben tener en cuenta las licencias de los todos los vocabularios conectados (Korn & Oppenheim, 2011; Rodríguez-Doncel et al., 2013).

Desde un punto de vista meramente práctico, la declaración de licencias en cualquier vocabulario de la familia RDF se efectúa mediante sus mismos elementos y esquemas. Un dataset puede licenciarse mediante declaración interna o mediante fichero adjunto en los metadatos VoID o DCAT, para ello puede utilizar las propiedades `void:Dataset`, `dcat:Dataset` o `dctype:Dataset`. En el caso de los mapeos las propiedades `void:Linkset` y `void:target` identifican el dataset referido y declaran la licencia mediante `dcterms:license`:

```
<http://www.upm.es/biblioteca/kos/sh#> a void:Linkset ;  
void:target <http://data.bnf.fr/ark:/> .  
<http://data.bnf.fr/ark:/> a void:Dataset .  
dcterms:license <http://creativecommons.org/publicdomain/zero/1.0/>.
```

En el ejemplo anterior, la propiedad `dcterms:license`, identifica la información legal que define la utilización del recurso, cabe también la utilización de `dcterms:rights`, que expresa los derechos sobre el contenido del recurso globalmente considerado y `dcterms:accessRights`, que expresa las condiciones de accesibilidad al recurso (Rodríguez-Doncel et al., 2013).

Una vez establecido que el contenido de datos vinculados puede ser protegido de diferentes maneras, cabe un nivel más granular de expresión de derechos que informe más ampliamente de la estructura legal que contienen los datos. Para ello existen vocabularios y ontologías definidas para profundizar en la declaración de derechos. Quizás las más representativas son:

1. ODRL: Open Digital Rights Language 2.0, Ontology (Iannella, Guth, Paehler, & Kasten, 2012): diseñada para la declaración de derechos de contenidos digitales en todos los sectores, ofreciendo un modelo abierto, que permite expresar tanto los derechos tradicionales, declaraciones Open Access e incluso expresiones de privacidad en redes sociales.
2. ccREL: The Creative Commons Rights Expression Language: especificación en RDF que ofrece una plataforma de expresión de licencias Creative Commons a creadores y publicadores de recursos digitales (Abelson, Addida, Linksvayer, & Yergler, 2008).
3. ODRS: Open Data Rights Statement Vocabulary: vocabulario de definición de derechos en Open Data, ofrece un lenguaje legible por máquina que permite declarar licencias, avisos de copyright y requisitos de atribución asociados a la reutilización y publicación de datasets (Dodds, 2013).
4. Linked Data Rights: es una extensión de la ontología ODRL que especifica elementos que permiten la declaración de los derechos de los datos en los recursos identificados como Linked Open Data. Este vocabulario está siendo desarrollado por el Ontology Engineering Group de la Universidad Politécnica de Madrid (Rodríguez, Poveda-Villalón, Suarez, & Gomez, 2013).

La inclusión de licencias correctas y claras de uso es un requisito fundamental del consumo que debería ser obligatorio. No son pocas las publicaciones que carecen de licencia de uso o que no la expresan correctamente, quizás por la falta de vocabularios específicos. También se deben desarrollar de modo más completo el sistema de licenciamiento de mapeos. En definitiva, se puede afirmar que los fundamentos de la calidad de los datos son la veracidad y la fiabilidad, la descripción de la procedencia y el adecuado licenciamiento (Korn & Oppenheim, 2011; Rodríguez-Doncel et al., 2013).

7.2.2 Marco jurídico general

La Directiva 2013/37/UE del Parlamento y del Consejo de la Unión Europea de 26 de junio de 2013 por la que se modifica la Directiva 2003/98/CE relativa a la reutilización de la información del sector público (Parlamento Europeo & Consejo de la Unión Europea, 2013), regula que sus organismos (aquellos que no están expresamente eximidos por la Directiva, y las bibliotecas y centros análogos ya no lo están) tienen la obligación respecto a la información que manejan en el ejercicio de sus actividades de (en aspectos referidos a licencias):

1. Ofrecer información reutilizable de modo no discriminatorio sin prohibición de su utilización no comercial.
2. Poner a disposición la información en formatos estándar y lenguaje común y cuando sea posible y proceda, en abierto y en lenguaje legible por máquina junto con sus metadatos.
3. Promover medidas que faciliten la búsqueda de la información para su reutilización.

La Directiva 2013/37/UE (Directiva PSI) indica que las licencias que se utilicen no deberían oponer restricciones a la reutilización de los datos. Las licencias deben estar disponibles en línea y ofrecer características de reutilización geográficas, tecnológicas y financieras. Los organismos del sector público podrán autorizar la reutilización sin condiciones, o con ellas, como por ejemplo reflejar la procedencia de los datos, aunque estas condiciones no deben restringir innecesariamente las posibilidades de reutilización y no se usarán para restringir la competencia (Parlamento Europeo & Consejo de la Unión Europea, 2013).

7.2.3 Asignación de la licencia al proyecto Tesoros Materias BUPM

La licencia escogida para el proyecto es la Attribution 4.0 International (CC BY 4.0). Esta licencia permite compartir el recurso bajo cualquier formato o medio. El reutilizador puede mezclar el contenido con otros trabajos, transformarlo o ampliarlo incluso con fines comerciales. Se prohíbe al reutilizador revocar estos derechos y se requiere la atribución del trabajo, ofreciendo el link de la licencia (`cc:attributionURL`). Se prohíbe la aplicación de cualquier tipo de restricción tecnológica o legal a la utilización de este trabajo tal y como queda configurado en esta licencia (Creative Commons, 2014a).

Representación del fichero de declaración de licencia del proyecto de tesoro de materias BUPM.

```
@prefix cc:      <http://creativecommons.org/ns#> .
@prefix dct:     <http://purl.org/dc/terms/> .
@prefix dc:      <http://purl.org/dc/elements/1.1/> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml:     <http://www.w3.org/XML/1998/namespace> .
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .

<http://academia6.poolparty.biz/PoolParty/wiki/Tesoro-Materias-BUPM>
a cc:Work ;
cc:license < http://creativecommons.org/licenses/by/4.0/> ;
cc:attributionName "Rafael Ávila"
cc:attributionURL <http://academia6.poolparty.biz/PoolParty/wiki/Tesoro-
Materias-BUPM> .
dc:format "text/turtle" ;
# http://www.iana.org/assignments/media-types/media-types.xhtml#
dct:title "Tesoro de materias BUPM"
dc:type <http://purl.org/dc/dcmitype/Dataset> .

<http://creativecommons.org/licenses/by/4.0/> a cc:License ;
cc:permits
    cc:DerivativeWorks,
    cc:Distribution,
    cc:Reproduction ;
cc:requires
    cc:Attribution,
    cc:Notice .
```

(Creative Commons, 2014b)

7.3 PRESERVACIÓN EN LINKED DATA

7.3.1 ASPECTOS GENERALES DE LA PRESERVACIÓN DE DATOS

La preservación digital es uno de los temas prioritarios en el desarrollo y continuidad de Linked Data. A modo de descripción filosófica podemos introducir la idea de preservación digital a través de las consideraciones de grupo de trabajo “Blue Ribbon” sobre preservación y acceso digital sostenible: “....*access to information tomorrow depends on preservation actions taken today. A fundamental fact of digital sustainability is that without preservation, there is no access.*”. (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010)

La preservación digital supone la utilización de soluciones tecnológicas para permitir el acceso a los datos en el futuro. Cualquier esfuerzo presente en la generación de datos vinculados debe estar acompañado de una planificación de la usabilidad futura de los mismos en óptimas condiciones. Para ello se deben establecer los procesos adecuados que permitan el mantenimiento del valor y calidad de los datos en cualquier utilización a lo largo del tiempo.

La teoría general de la preservación refiere que son varias las cuestiones que afectan principalmente al mantenimiento del acceso a los datos a lo largo del tiempo:

1. El carácter cambiante de la tecnología que afecta tanto a los datos (obsolescencia de formatos de archivo) como a las aplicaciones informáticas que los gestionan.
2. El envejecimiento (u obsolescencia) de los soportes informáticos que se usan como contenedores de los datos.
3. Fallo crítico de los sistemas de almacenamiento.
4. La imposibilidad para poder determinar el valor o la utilidad de la información digital, por ejemplo por la pérdida de la documentación que lo refería.
5. La pérdida del objeto mismo o información digital de modo irreversible.

Desde esta perspectiva, las estrategias de preservación deben corresponderse con las mejores prácticas al efecto

1. Planes tecnológicos para la preservación.
2. Estrategias de emulación de tecnologías.
3. Estrategias de migración de recursos, formatos y tecnologías.

(Antoniou et al., 2014a; Antoniou et al., 2014b).

No podemos profundizar aquí en un tema tan complejo y de tanto calado como el mantenimiento de los datos y su valor a lo largo del tiempo, pero podemos apuntar algunas pautas recomendadas para la preservación de los datos:

1. Protección física de equipos y soportes, preservación de datos por duplicación, preservación por establecimiento de copias de seguridad en diferentes ubicaciones, preservación de ficheros mediante migraciones a nuevos formatos, preservación de sistemas y aplicaciones, preservación por mantenimiento de licencias, preservación de los entornos de representación de datos y preservación de la documentación.
2. Utilizar formatos de archivo en estándares abiertos.
3. Utilizar los servicios de los repositorios digitales para la preservación.
4. Crear y mantener documentados todos los procesos implicados en la preservación a través de metadatos (PREMIS).
5. Emplear estrategias avanzadas de almacenado distribuido para proteger los recursos digitales.

PREMIS (Preservation Metadata Implementation Strategies) es un producto para la preservación compuesto de un diccionario de datos que se emplean junto a los metadatos de preservación de recursos digitales (según el propio diccionario PREMIS, los metadatos de preservación son la propia información, de cualquier tipo, que emplea un repositorio digital para establecer la preservación). Estos datos estructuran a los metadatos de preservación en cada una de las fases de la misma. (Library of Congress, 2014c).

7.3.2 PRESEVACIÓN DE DATOS VINCULADOS

En el contexto Linked Data, la preservación añade nuevos retos a los clásicos de la preservación digital, como es el sostenimiento de la calidad y usabilidad en la estructura de links, la continua evolución de los formatos de estructuración de datos, la heterogeneidad de sus tipos y calidad de las publicaciones, variabilidad de vocabularios, etc. La tendencia en la preservación en Linked Data viene de la mano de tres puntos fundamentales: la mejora de la descripción de los datasets, la definición de las sucesivas versiones de los mismos y el empaquetamiento de los ficheros que componen el conjunto de la publicación (Cyganiak, 2013).

El mantenimiento de las estructuras de vinculado es una tarea compleja. La integridad del sistema de enlaces a través de IRIs es un elemento primordial para la preservación en Linked Data. Evitar la proliferación de enlaces rotos es un objetivo necesario y difícil de conseguir, ya que por su propia naturaleza, la Web es un entorno cambiante que inevitablemente rompe las referencias entre recursos. Por ello, más que trabajar por la integridad de los IRI, puede ser más eficaz la creación de sistemas de auto reparación de los links entre datos. Existen procesos que permiten aprovechar el modelo de datos vinculados para recuperar datos perdidos por enlaces rotos. Se

trata de algoritmos que utilizan los IRIs y las estructuras de enlace para “reconstruir” la información que se ha podido perder a través de la información que el propio IRI nos ofrece. Se establece pues una estructura de preservación a posteriori, solucionando problemas de rotura de enlaces, por otra parte casi inevitables. (Vesse, Hall, & Carr, 2010) Por todo ello, y entre otras actividades, se está focalizando la atención en los “*RDF data dumps*” (los ficheros descargables de datos vinculados), lo que permite un mejor procesado conjunto de los datos para su preservación (Antoniou et al., 2014b; Cyganiak, 2013).

La preservación debe abarcar el mantenimiento de vocabularios o la definición de sus sucesivas versiones. Aun así, los principales registros de vocabularios, LOV y OMR, (como ya se dijo en el capítulo anterior) disponen de sistemas de actualización y control de versiones. Claro ejemplo de ello es la eficacia con la que se ha dispuesto en OMR las actualizaciones de los *namespaces* en RDA. El versionado para la preservación requiere nuevas técnicas que simplifiquen el procesado automático de los datasets. Para ello se han definido parámetros y procesos para el empaquetado de los datasets y la información descriptiva o técnica sobre ellos. Open Knowledge Foundation ha dispuesto un borrador de procesos de empaquetamiento que está todavía en fase de desarrollo. Igualmente importante es definir las dependencias entre los datos vinculados y aquellos vocabularios u otros datasets con los que estén vinculados. Otra de las tendencias actuales es la incorporación de procesos de preservación provenientes del estándar para la conservación de archivos Open Archival Information System (OAIS), cuyo modelo ha sido desarrollado por la comunidad de datos digitales en el contexto de los archivos y que permite utilizar una serie de procesos y principios de conservación de archivos para la preservación a largo plazo de los datos y por extensión de los vinculados. Uno de los componentes más eficientes y novedosos para conservar la calidad de los datos son los Trusted Digital Repositories (TDR), que permiten de modo estructurado y formal, la gestión de datos de calidad, sometiéndose a procesos estandarizados de auditoría y certificación. (Antoniou et al., 2014a; Antoniou et al., 2014b; Giaretta, 2011).

7.3.3 PRELIDA

El proyecto PRELIDA pretende poner en contacto a las comunidades de la Web Semántica y de la preservación digital, para un apoyo recíproco a la hora de gestionar la preservación de datos. Se trata también de que no se desarrollen dos vías separadas en cuanto a la preservación de datos digitales y por ello se tiende a unir sus desarrollos en cuanto a preservación digital y preservación en Linked Data sobre todo en el contexto GLAM.

El proyecto parte del firme convencimiento de que ambas comunidades se necesitan y se complementan. Por una parte se debe mostrar a la comunidad Linked Data que hay un marco muy desarrollado de preservación para recursos digitales, del cual se puede extraer buenas prácticas para preservar los datos vinculados. Por otro lado la comunidad de la preservación

digital debe conocer las especificaciones y retos que supone la conservación a largo plazo de datos vinculados: su estructura distribuida, su capacidad de vinculación mediante IRIs, el entorno en constante cambio en el que se desenvuelve (Seventh Framework Programme FP7/2013, 2014).

Para ello, el proyecto promueve una serie de actividades tendentes a respaldar su objetivo de hacer evolucionar la preservación digital en el contexto Linked Data:

1. Reunir, organizar y publicar casos de uso especialmente relevantes para la preservación digital en Linked Data.
2. Crear un corpus de buenas prácticas a través del estudio del estado del arte en estos temas.
3. Establecer la vigilancia tecnológica al respecto.
4. Reunir a las partes implicadas en el estudio de los retos presentes y futuros, promoviendo la preparación de estrategias para adecuarse a los cambios.
5. Promover la estandarización de procesos de preservación entre los organismos pertinentes.

7.3.3.1 PROCESOS DE PRESERVACIÓN DEL PROYECTO SEGÚN RECOMENDACIONES PRELIMINARES

Se distinguen dos tipos de publicaciones principales de datos vinculados a efectos de preservación. Publicación puramente web de los recursos semánticos y publicación en “triple-stores” a modo de bases de datos.

La preservación de los recursos en Linked Data en la Web se realiza de modo análogo a la preservación de páginas web, incluso con más facilidad si los datos internos de la web están semánticamente referenciados a través de enlaces persistentes como los IRIs. Para las publicaciones web, los principales problemas son los links que dejan de funcionar y la obsolescencia del contenido vinculado. A parte de la estrategia algorítmica de recuperación de contenido descrita más arriba, existe la posibilidad de luchar contra los enlaces rotos a través del establecimiento de versiones identificadas por etiquetas temporales en los IRIs, lo que permite proteger la recuperabilidad acudiendo a versiones anteriores etiquetadas. Para los datos vinculados almacenados en bases de datos específicas, las tareas de preservación se asemejan a las de cualquier base de datos (Antoniou et al., 2014b; Cyganiak, 2013; Seventh Framework Programme FP7/2013, 2014).

El dataset descrito en este proyecto reside en un “triplestore” (Semantic Web) bajo tecnologías propietarias y sin control de tareas de preservación por parte de los editores. Aparte de la confianza en las infraestructuras utilizadas por terceras partes, se debe analizar las posibles situaciones que este contexto genera:

1. La simple preservación de los datos brutos no genera problemas. Como datos estáticos, su preservación debe seguir las pautas de la disciplina de preservación general: formatos no propietarios que faciliten la migración, almacenamiento distribuido y control sobre las aplicaciones que permiten trabajar con esos formatos (también los sistemas que dan soporte a las aplicaciones). En el contexto de este proyecto los datos y su estructura serán publicados en diferentes formatos y serializaciones. A efectos de compatibilidad con la plataforma se utilizará el formato de serialización TRIG, que permite conservar toda la información aneja al dataset y que permitiría un posible reemplazo del tesoro ante fallos en el servidor. Se establecerán otras serializaciones en Turtle, JSON-LD y en el formato nativo de la base de datos.
2. Las diversas serializaciones deben mantenerse y migrar a otras nuevas para mantener la compatibilidad en su caso. Utilidades de programación pueden ser necesarias para transformar dichos formatos en otros nuevos. Dichas utilidades deben ser preservadas y sometidas al control de un plan de vigilancia tecnológica que verifique los cambios en todas las tecnologías utilizadas.
3. Como objeto digital dinámico, un dataset estructurado en Linked Data y sometido a constantes revisiones y actualizaciones como el que tratamos aquí, nos obliga a adoptar políticas de preservación de versiones que definan cuales mantener y con qué criterio. El sistema gestor de tesoros permite exportaciones diversas que pueden ser descritas con metadatos de preservación conteniendo etiquetas temporales y descripción de los cambios. Estas versiones exportadas son producto de preservación y fácilmente recuperables por importación. Como parte del propio proceso del proyecto se está siguiendo un protocolo de backup diario de las adiciones al tesoro en formato Trig y mediante etiquetas temporales.
4. Los datos brutos con los que se han creado los vocabularios de materias no tienen definida una estructura semántica. Ésta se construye mediante el sistema de relaciones y de control de vocabulario que nos dirigen a precisar una construcción original que debe ser preservada. Las utilidades de exportación permiten extraer datos modelados con etiquetas de definición de relaciones en los tesoros (BT, NT, RT) o extraer los datos con la semántica definida mediante los vocabularios para la definición de KOS. Preservar la primera estructura puede ser eficaz para migrar a nuevos vocabularios para la organización del conocimiento. Preservar la segunda, nos permite obtener una foto fija de una descripción semántica compleja y temporalmente definida, que puede migrarse por programación, recogiendo toda la riqueza descriptiva del registro de materias. La

preservación de versiones es aquí fundamental. Se trata pues de preservar el conocimiento contenido en las declaraciones semánticas.

5. No es posible gestionar la preservación de los vínculos ofrecidos por terceros. El utilizar servidores y servicios de terceros tanto para el almacenamiento como para la gestión de IRIs, puede provocar problemas de acceso futuro a los datos. Por ello se debe gestionar una total migración a infraestructuras propias que permitan tener control sobre todos los aspectos relacionados con la preservación y en concreto con los sistemas de vinculación o IRIs.
6. Para facilitar el proceso de acceso automático y la preservación se ofrecerán en el contexto de este proyecto conjuntos empaquetados para facilitar la identificación de versiones, según proyecto *Datapackage* de OKFN.
7. Los mapeos con otros vocabularios pueden ser preservados mediante un servicio de redirecciones que debe establecerse en los propios servidores con enrutados alternativos de las peticiones a los servidores ajenos. En cambio preservar la calidad semántica de la relación no es posible de modo general si se producen cambios importantes en los vocabularios objetivos. La preservación en un contexto de interoperabilidad no es un tema resuelto.
8. Cualquier procesado de los datos que se establezca, supondrá la utilización de tecnologías diversas que deben preservarse e incluirse en el inventario de la vigilancia tecnológica.
9. La evolución de los diferentes vocabularios que describen los datos no parece un gran problema de preservación estableciendo migraciones programadas y automatizadas.
10. Cualquier información deber ser preservada, también las licencias y sus formatos, los metadatos, información para la utilización del dataset y en general cualquier información complementaria necesaria para el mantenimiento del uso de los datos.
11. Desde el punto de vista de la preservación local se establecerán controles de versiones del dataset (Se incluye fichero “Provenance” describiendo las diferentes etapas de la construcción del vocabulario hasta su publicación), aseguramiento de los soportes de datos, respaldo de publicaciones, diseminación geográfica de copias, migración en lo posible a formatos no propietarios, evaluaciones de control de calidad y establecimiento de estrategias de migración, de emulación o de cambio tecnológico completo. Estrategias de migración actualización y vigilancia tecnológica para prever los cambios de

tecnologías: cambios de formatos, vocabularios, técnicas de enlace, técnicas de publicación, testeo de validez.

El sistema de empaquetado que vamos a utilizar requiere establecer un sistema de descripción o catalogación semántica mediante JSON-LD, que contendrá metadatos sobre los metadatos y los recursos tal y como se puede ver en la imagen siguiente (Open Knowledge Foundation, 2014b).

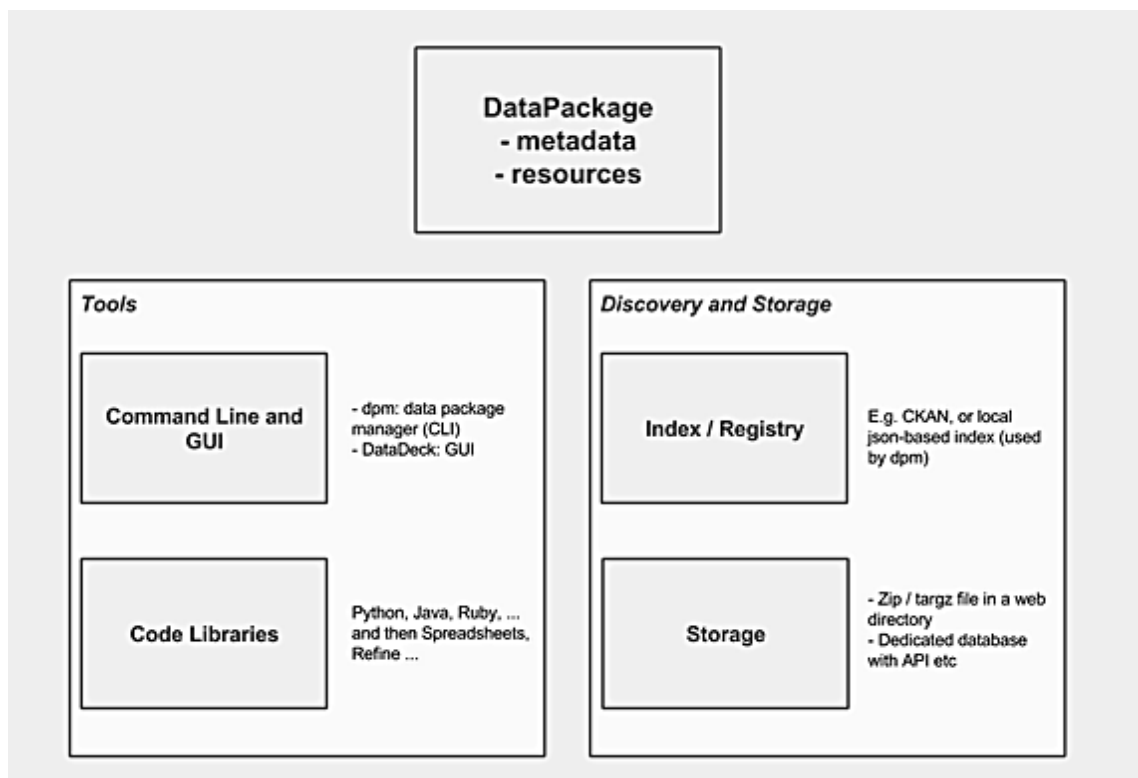


Figura 32 Estrategia de empaquetado para la preservación. Fuente: OKFN 2014

7.3.3.2 PROPUESTA DE EMPAQUETADO DEL SISTEMA DE FICHEROS DEL VOCABULARIO BUPM

Las publicaciones Linked Open Data al completo se componen de varios ficheros con diferentes funciones: los propios datos con su específica estructura, fichero de licencias, ficheros de metadatos, archivos de ayuda o información complementaria, links que acompañan a los dataset, etc. A efectos de visibilidad hemos recomendado la publicación por separado de esta información, lo que permite discriminar los diferentes aspectos de una publicación pero a la vez complica la preservación al existir varios objetos y diferentes formatos que proteger. La solución para ello es establecer una estructura de empaquetado de ficheros (técnica que proviene de la asociación de

archivos en contextos de programación) que unifique en lenguaje máquina los diferentes productos de una publicación. La técnica de empaquetado de datos que propone Open Knowledge Foundation no es un estándar, más bien es un proceso todavía en fase de asentamiento y de aceptación por la comunidad de la Web de datos.

Un empaquetado de datos se compone de metadatos sobre los recursos describiendo su estructura y contenido y opcionalmente ficheros adicionales. Un empaquetado se plasma en un fichero en JSON que identifica con metadatos los recursos de la publicación. Este fichero debe estar emplazado en el directorio raíz del recurso. Los tipos más habituales descritos en el empaquetado son ficheros locales, ficheros remotos y sus URLs y recursos en línea asociados a los recursos.

Contenido del fichero datapackage.json

1. Metadatos generales.
2. Listados de recursos incluidos en el empaquetado
3. Información adicional.

```
{
  "name": "Tesauro-Materias-BUPM",
  "title": "Tesauro de materias de la Biblioteca de la Universidad Politécnica de Madrid",
  "description": "Lista de encabezamientos de materia de la biblioteca de la Universidad Politécnica de Madrid publicados en Linked Open Data",
  "homepage": " http://academia6.poolparty.biz/PoolParty/wiki/Tesauro-Materias-BUPM/",
  "version": "1.0",
  "licenses": [{
    "type": "CC-BY-4.0",
    "url": "http://creativecommons.org/licenses/by/4.0/"
  }],

  "sources": [{
    "name": "Tesauro de materias BUPM",
    "web": " http://academia6.poolparty.biz/PoolParty/wiki/Tesauro-Materias-BUPM/"
  }],
  "keywords": "encabezamientos de materias, vocabularios semánticos, library Linked data",
  "datapackage_version": "1.0",

  "author": [ {
    "name": "Rafael Ávila",
    "email": "rafael.avila@mail.com",
    "web": "http://www.example.com"
  } ],

  "temporal": "07/20/2014;",

  "contributors": [ {
    "name": "Virginia Ortíz Repiso",
    "email": "vrepiso@mail.com",
```

```

    "web": "http://www.example.com"
  }],

  "resources": [{
    "url": "https://ckannet-storage.commondatastorage.googleapis.com/2014-07-27T21:32:55.672Z/tesauro-bupm.trig",
    "name": "Tesauro-bupm",
    "type": "dataset",
    "description": "Fichero en TRIG que contiene la lista de encabezamientos de materia BUPM estructurada en SKOS."
  },
  {
    "url": "https://ckannet-storage.commondatastorage.googleapis.com/2014-07-28T10:28:42.182Z/void-bupm-sh.ttl",
    "name": "VoID-BUPM",
    "format": "turtle",
    "mediatype": "text/turtle",
    "description": "Fichero en Void de asignación de metadatos descriptivos al Tesauro de materias BUPM"
  },
  {
    "url": "http://academia6.poolparty.biz/PoolParty/sparql/Tesauro-Materias-BUPM",
    "name": "SPARQL Endpoint",
    "type": "online API",
    "description": "Punto de acceso al cliente SPARQL del Tesauro de Materias BUPM"
  },
  {
    "url": "https://ckannet-storage.commondatastorage.googleapis.com/2014-07-28T16:01:28.523Z/dcat-bupm.ttl",
    "name": "DCAT-BUPM",
    "mediatype": "text/turtle",
    "description": "Fichero DCAT para la descripción del dataset del Tesauro de materias BUPM en el contexto de catálogos de datos"
  },
  {
    "url": "https://ckannet-storage.commondatastorage.googleapis.com/2014-07-28T15:52:04.210Z/provenance-bupm.ttl",
    "name": "Provenance-BUPM",
    "type": "text/turtle",
    "description": "Descripción con metadatos Provenance del proceso de generación del Tesauro de materias BUPM"
  },
  {
    "url": "https://ckannet-storage.commondatastorage.googleapis.com/2014-07-28T16:58:36.031Z/licencia-ccby40-bupm.ttl",
    "name": "Licencia CC BY 4.0",
    "type": "text/turtle",
    "description": "Asignación de licencia Creative Commons Attribution 4.0 International (CC BY 4.0) al Tesauro de materias BUPM"
  }
]
}

```

Verificado con <http://json.parser.online.fr/>

8 CONCLUSIONES

Dos han sido los objetivos fundamentales fijados para concluir con éxito este trabajo: evolucionar desde un vocabulario de materias, sistematizado en una base de datos de uso local, hacia formas de organización del conocimiento más adaptadas al entorno digital, y por otra parte, la descripción mediante técnicas semánticas de ese producto para propiciar su reutilización, convirtiéndolo en un vocabulario interoperable, multilingüe y compatible libremente.

Para conseguirlo se ha efectuado un acercamiento a la literatura más reciente, desde la más contextual a la más orientada al objeto de este trabajo, todo ello para conseguir una adecuada visión de las posibilidades que ofrecía el estado actual de las disciplinas implicadas.

El primer reto, la migración del vocabulario, ha tenido una clara conclusión general. La construcción o modificación de un vocabulario debe contar necesariamente con un completo equipo multidisciplinar, bien organizado, que afronte las diferentes aristas que se vayan a presentar en su desarrollo. No por encontrarnos en un mundo tecnológicamente complejo, las tareas de gestión de vocabularios dejan de ser esencialmente intelectuales, y especialmente necesarias para la correcta estructuración semántica de las relaciones. Sin un conocimiento profundo de los campos semánticos, no es posible establecer altos niveles de calidad en la sistematización del vocabulario. Si además se añade, que percibir el significado de los conceptos es especialmente complejo si se aborda el establecimiento de mapeos con otros vocabularios en otros idiomas, el círculo de dificultades se completa.

Aunque estas dificultades las tuviéramos superadas, la plasmación de nuestro trabajo en herramientas de gestión de vocabularios de tipo general no es suficientemente satisfactoria. Ya se dijo que la técnica más adecuada es contar con expertos en sistemas que proporcionen herramientas que se adapten, al menos, a los requerimientos básicos de expresión del vocabulario, so pena de tener una dependencia peligrosa, no sólo respecto a la limitada expresividad semántica, sino también en relación a la gestión de vínculos, descripción de metadatos o posibilidades de ofrecer una estrategia válida de preservación. Además, las mínimas condiciones de fiabilidad y calidad obligan a emplear medios automáticos para efectuar los modelados, pues la alternativa manual no es eficiente e introduce la posibilidad del error en porcentajes más altos que lo deseable.

A pesar de los obstáculos, los resultados obtenidos corroboran algunos aspectos positivos ya conocidos. La calidad de los vocabularios de materias, cuya estructuración ha servido de ejemplo para solucionar no pocos problemas de gestión y mapeo. Las ventajas que ofrece la disponibilidad en abierto y las posibilidades de reutilización de los vocabularios que permite vincularlos y establecer plataformas multilingües. La disponibilidad de herramientas para la gestión de vocabularios, que a pesar de su complejidad, permiten la realización práctica de los objetivos fundamentales. Y finalmente, la posibilidad de contar con normas actualizadas y estandarizadas

para el contexto digital, que guían de modo relativamente sencillo hacia la producción completa de un vocabulario. Es cierto que la norma ISO 25964 1 y 2 se orientan más a las aplicaciones que a la Web de datos, pero cuando las tareas de alineamiento de SKOS e ISO 25964, estén finalizadas, la gestión y expresión semántica de vocabularios será probablemente una tarea mucho más sencilla.

Como se ha dicho, otro objetivo nuclear ha sido la migración a un modelo semántico de representación del vocabulario mediante técnicas y tecnologías Linked Data. Aquí lo más destacable es la posibilidad de utilización de un estándar como SKOS, que no por tener una implementación básica y generalista, deja de ser una herramienta valiosísima y estable para la representación de KOS. Sus carencias, como se ha visto, residen más en conseguir un esquema lo más sencillo posible, que permita que cualquier vocabulario representado sea automáticamente interoperable con los demás. Habría que valorar si una mayor complejidad de SKOS mediante extensiones, no redundaría en un aumento de complejidad que penalizara su utilización generalizada.

La inclusión del modelo para etiquetas SKOS-XL abre la posibilidad a un enriquecimiento importante en el modelado a nivel léxico, incrementando la expresividad y posibilidades de relación de conceptos y sus términos. La utilización de ontologías o modelos unilaterales no creo que sea el camino para extender la producción semántica en la publicación de vocabularios de calidad. La excesiva complejidad que añaden y las dificultades de interoperabilidad que pueden suponer, debería desincentivar su utilización. El problema es decidir qué hacer: ¿establecer ontologías propias y locales?, ¿desarrollar lenguajes como MADS?, ¿extender SKOS para ampliar su ámbito de representatividad? o ¿mantener niveles más ágiles y sencillos pero con menor granularidad en la representación de KOS?.

Como se ha visto, en el contenido de este trabajo se han intentado ofrecer modos de representación de los aspectos más complejos del vocabulario, como la precoordinación, el modelado de colecciones, las polijerárquicas o el mapeo en supuestos de equivalencia compuesta. Recoger toda la profundidad del contenido semántico de los vocabularios es correcto, pero ¿es imprescindible? La simplicidad de los componentes del ecosistema de la Web ha sido el substrato de su éxito. A mi juicio, simplificar la gestión semántica de vocabularios supondría probablemente la generalización de la publicación de vocabularios bibliotecarios de calidad en Linked Data.

Las herramientas de trabajo han sido otro de los aspectos complejos que ha habido que afrontar. En general las aplicaciones para el trabajo semántico no son fáciles ni de instalar ni de manejar, se alejan de los estándares de usabilidad que permitirían un acercamiento más atrevido a estas tecnologías. De todos modos, también hay que decir que se perciben importantes avances en este tema, el ejemplo claro es la aplicación empleada para la gestión del tesoro y su publicación en LOD, que ha demostrado que a pesar de sus carencias (la más importante es no poder establecer mapeos con otros vocabularios que los predefinidos) ofrecen un rango de utilización y de calidad de buen nivel.

En otro orden de cosas, es importante la proliferación de iniciativas de toda índole en la transformación de productos de la información en su versión semántica. Este despegue de los datos vinculados indica que podemos estar en el punto crítico a partir del cual la producción de representaciones semánticas aumente progresivamente. Ejemplo claro son los datos de investigación que, gracias al apoyo económico de la UE, abrirán un campo de trabajo con amplia disponibilidad de recursos, requiriendo sin duda, la utilización de tecnologías semánticas de tratamiento de la información.

Siendo esto así, se echa de menos una mayor compenetración de esfuerzos. Aunque se está realizando un gran esfuerzo normalizador y se trabaja en la estandarización de productos y procesos, la realidad parece indicar que estamos ante desarrollos en paralelo pero con escaso nivel de interacción entre los diferentes agentes productores. En el caso de los vocabularios de materias, no existe coordinación entre los diferentes productos de referencia. Las bibliotecas de cabecera del sistema crean sus propios métodos de representación, fundamentados en SKOS sí, pero complementados con lenguajes y procedimientos específicos de muy diferente índole. Hubiera sido deseable, un desarrollo apoyado en el primer vocabulario disponible en LOD, las LCSH, haciendo evolucionar los subsiguientes proyectos desde ellas y generando un proceso conjunto. No se trata de dar preferencia a uno de los vocabularios, se trata de evolucionar y mejorar lo que se tiene pero desde una estructura común. Ciertamente es que la interoperabilidad soluciona algunos de estos problemas como se ha visto, pero también es cierto que es un recurso potente pero limitado que no puede servir para vincular cualquier tipo de estructura. Podríamos hablar de la necesidad de establecer una interoperabilidad ya no de los datos, sino de los servicios de información, en un contexto de interacción de datos la vinculación de bibliotecas debería ser el modo escogido de trabajo.

A nivel local el problema es el mismo, debería de existir un origen común en el despliegue de tecnologías semánticas, la biblioteca cabecera de sistema establecería las pautas de coordinación con el resto de agentes (las universidades por ejemplo), desarrollando productos únicos con extensiones locales tanto en el ámbito de los vocabularios, los modelos para representar contenidos o las políticas de representación de conjuntos de datos.

Desde este trabajo se ha defendido la necesidad de publicar en LOD los productos de las bibliotecas como los vocabularios de valores. Su calidad puede ser una herramienta válida para la indización y la recuperación de los recursos digitales. Llevada a cabo la tarea de cohesión antes referida, la publicación de diferentes vocabularios de materias alineados y en diferentes idiomas, sería una potente herramienta para vertebrar el conocimiento en la web. La creación de un sistema de gestión y representación de materias multilingüe, al nivel de VIAF, podría aportar grandes ventajas en la asignación de metadatos de calidad, y la recuperación de grupos temáticos en diferentes idiomas, aunque las búsquedas se efectuaran en uno sólo de ellos. Tecnología suficiente hay para ello, hace falta voluntad para acometer un proyecto de esa índole.

Los vocabularios de valores debido a la calidad de alguno de ellos, son unos candidatos idóneos para su utilización como datos vinculados. No es cierto que los más significados productos de la

información tradicionales ya no sean útiles; en realidad en este trabajo se ha constatado que el trabajo intelectual es el más importante pues, lo verdaderamente complicado es generar un buen vocabulario de valores o una asignación de metadatos correcta para describir contenidos, no su presentación en LOD; esas competencias siguen siendo necesarias. En definitiva, los servicios de información deben de implicarse de modo fundamental y coordinado con la representación del conocimiento digital (o quizás otros lo hagan por ellos). Las tecnologías semánticas y en concreto Linked Open Data, puede ser útiles para hacerlo, pero no debe ser considerada la única vía posible, el objetivo es representar el conocimiento, no utilizar Linked Data. Además la simplicidad debe ser el contexto elegido para conseguir el objetivo, sin ella el fracaso es más que probable. Las ventajas pueden ser insospechadamente importantes, una evolución en la gestión del conocimiento sin precedentes e importantes beneficios para la Sociedad, además de la no menos importante, volver a colocar a las bibliotecas en una posición protagonista en el nuevo espacio de la información.

9 REFERENCIAS

- Abelson, H., Addida, B., Linksvayer, M. & Yergler, N. (2008). ccREL: The creative commons rights expression language. Retrieved from <http://www.w3.org/Submission/ccREL/>
- American Library Association. (2014). RDA toolkit. resources description & access. Retrieved from <http://www.rdatoolkit.org/>
- Antoniou, G., Batsakis, S., Isaac, A., Scharnhorst, A., García, J. M., Van Horik, R., . . . Meghini, C. (2014a). *Analysis of the limitations of digital preservation solutions for preserving LD*. (Deliverable No. 600663).PRELIDA.
- Antoniou, G., Batsakis, S., Isaac, A., Scharnhorst, A., García, J. M., Van Horik, R., . . . Meghini, C. (2014b). *State of the art assessment on linked data and digital preservation*. (Deliverable No. 600663).PRELIDA.
- Archer, P., Goedetier, S., & Loutas, N. (2012). *Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC*. ().PwC EU Services. Retrieved from https://joinup.ec.europa.eu/sites/default/files/D7.1.3%20-%20Study%20on%20persistent%20URIs_4.pdf
- Ashley, K. (2012). Research data and libraries: Who does what. *Insights*, 25(2), 155-157. doi:10.1629/2048-7754.25.2.155
- Auer, S., Lehmann, J., Ngonga Ngomo, A., & Zaveri, A. (2013). Introduction to linked data and its lifecycle on the web. In R. Rudolph, G. Gottlob, I. Horrocks & F. van Harmelen (Eds.), *Reasoning web: Semantic technologies for intelligent data access 9th international summer school 2013 mannheim, germany, july 30 – august 2, 2013* [Sebastian Rudolph Georg Gottlob Ian Horrocks Frank van Harmelen] (pp. 1). Berlín: Springer. doi:10.1007/978-3-642-39784-4
- Ávila, R. (2014a). Página wiki de acceso al tesaurus de materias BUPM. Retrieved from <http://academia6.poolparty.biz/PoolParty/wiki/Tesaurus-Materias-BUPM>
- Ávila, R. (2014b). Representación del concepto "circuitos integrados" del tesaurus de materias BUPM. Retrieved from <http://academia6.poolparty.biz/PoolParty/wiki/Tesaurus-Materias-BUPM?URI=http://academia6.poolparty.biz/Tesaurus-Materias-BUPM/322>

- Ávila, R. (2014c). Representación visual del concepto "circuitos integrados" del tesauro de materias BUPM. Retrieved from <http://academia6.poolparty.biz/Tesauro-Materias-BUPM/322.visual>
- Baker, T., Bermés, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., . . . Zeng, M. (2011). Library linked data incubator group final report. Retrieved from <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/>
- Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). Key choices in the design of simple knowledge organization system (SKOS). *Web Semantics: Science, Services and Agents on the World Wide Web*, 20, 35-49. doi:10.1016/j.websem.2013.05.001
- Ball, A., & Duke, M. (2012). How to cite datasets and link to publications. DCC how-to guides. Retrieved from <http://www.dcc.ac.uk/resources/how-guides>
- Bauer, F., & Kaltenböck, M. (2012). *Linked open data: The essentials*. Viena: Semantic Web Company. Retrieved from <http://www.semantic-web.at/LOD-TheEssentials.pdf>
- Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., . . . Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2), 599-611. doi:10.1016/j.future.2011.08.004
- Beckett, D. (2014). RDF 1.1 N-triples. Retrieved from <http://www.w3.org/TR/2014/REC-n-triples-20140225/>
- Beckett, D., Berners-Lee, T., Prud'hommeaux, E. & Carothers, G. (2014). RDF 1.1 turtle terse RDF triple language W3C recommendation. Retrieved from <http://www.w3.org/TR/2014/REC-turtle-20140225/>
- Bermés, E. (2013). Enabling your catalogue for the semantic web. In S. Chambers (Ed.), *Catalogue 2.0: The future of library catalogue* (pp. 117-142). London: facet publishing.
- Berners-Lee, T. (2009). Linked data - design issues. Retrieved from <http://www.w3.org/DesignIssues/LinkedData.html>
- Biblioteca Nacional de España. (2014a). Datos BNE 2.0. Retrieved from <http://datos.bne.es/vocab>
- Biblioteca Nacional de España. (2014b). Datos enlazados en la BNE. Retrieved from <http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/index.html>
- Biblioteca Nacional de España. (2014c). Manual de autoridades. Retrieved from <http://www.bne.es/es/Micrositios/Guias/ManualAutoridades/Registros/EncabezamientosMateria/>

- Biblioteca Nacional de España. (2014d). Modelos (ontologías). Retrieved from <http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/Modelos/>
- Biblioteca Universidad Politécnica de Madrid. (2010). *Pautas para la gestión de registros de autoridad*. Unpublished manuscript.
- Bibliothèque nationale de France. (2014a). Bibliothèque nationale de france. Retrieved from <http://www.bnf.fr/fr/acc/x.accueil.html>
- Bibliothèque nationale de France. (2014b). RAMEAU subject headings as SKOS linked data. Retrieved from <http://www.cs.vu.nl/STITCH/rameau/>
- Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2011). The meaningful use of big data: Four perspectives – four challenges. *SIGMOD Record*, 40(4), 56-60. Retrieved from <http://www.sigmod.org/publications/sigmod-record/1112/pdfs/10.report.bizer.pdf>
- Bizer, C., & Cyganiak, R. (2014). RDF 1.1 TriG RDF dataset language W3C recommendation. Retrieved from <http://www.w3.org/TR/2013/WD-trig-20130409/>
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22. doi:10.4018/jswis.2009081901
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). *Sustainable economics for a digital planet ensuring long-term access to digital information*. ().Blue Ribbon Task Force on Sustainable Digital Preservation and Access.
- Breeding, M. (2014). Library resource discovery products: Context, library perspectives, and vendor positions. *Library Technology Reports*, 1, 1-58. doi:10.5860/ltr.50n1
- Brickley, D., & Guha, R. V. (2014a). RDF schema 1.1 W3C recommendation. Retrieved from <http://www.w3.org/TR/rdf-schema/>
- Brickley, D., & Guha, R. V. (2014b). RDF schema 1.1 W3C recommendation. Retrieved from <http://www.w3.org/TR/rdf-schema/>
- British Library. (2014). Welcome to bnb.data.bl.uk - the british national bibliography. Retrieved from <http://bnb.data.bl.uk/>
- Buchholtz, S., Bukowski, M., & Śniegocki, A. (2014). In Warsaw Institute for Economic Studies (Ed.), *Big and open data in europe: A growth engine or a missed opportunity?*. Warsaw: demosEUROPA. Retrieved from http://www.bigopendata.eu/wp-content/uploads/2014/01/bod_europe_2020_full_report_singlepage.pdf

- Carothers, G. (2014). RDF 1.1 N-quads: A line-based syntax for RDF datasets. Retrieved from <http://www.w3.org/TR/2014/PR-n-quads-20140109/>
- Cobden, M., Black, J., Gibbins, N., Carr, L., & Shadbolt, N. (2011). A research agenda for linked closed data. *Second international workshop on consuming linked data* (). Bonn: Retrieved from <http://eprints.soton.ac.uk/272711/3/position.pdf>
- Corson-Rikert, J., & Hidalgo, M. (2014). VIVO. Retrieved from <https://wiki.duraspace.org/pages/viewpage.action?pageId=34662505>
- Coyle, K. (2013). Linked data: An evolution. *Italian Journal of Library and Information Science*, 4(1), 53-62. Retrieved from <http://leo.cilea.it/index.php/jlis/article/view/5443/7889>
- Creative Commons. (2014a). Creative commons. Retrieved from http://creativecommons.org/choose/?lang=es_ES
- Creative Commons. (2014b). Creative commons namespace. Retrieved from <http://creativecommons.org/ns>
- Crupi, G. (2013). Beyond the pillars of hercules: Linked data and cultural heritage. *Italian Journal of Library and Information Science*, 4(1), 25-49. Retrieved from <http://leo.cilea.it/index.php/jlis/article/download/8587/7887>
- CSIC. (2012). Buenas prácticas y directrices para datos de investigación en digital.CSIC. Retrieved from <http://digital.csic.es/politicas/politicaDatos.jsp>
- Cyganiak, R. (2013). A perspective on preservation of linked data. In PRELIDA (Ed.), *1st PRELIDA workshop* ()
- Cyganiak, R., & et al. (2011). Guidelines for collecting metadata on linked datasets in CKAN - W3C wiki. Retrieved from http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKAN_metainformation
- Cyganiak, R., & Sauermann, L. (2008). Cool uris for the semantic web. W3C interest group note. Retrieved from <http://www.w3.org/TR/cooluris/>
- Cyganiak, R., Wood, D. & Lanthaler, M. (2014). RDF 1.1 concepts and abstract syntax W3C recommendation. Retrieved from <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#managing-graphs>
- Cyganiak, R., Zhao, J., Alexander, K. & Hausenblas, H. (2010). VoID guide - using the vocabulary of interlinked datasets. Retrieved from <http://vocab.deri.ie/void/guide/2009-01-29>

- Danskin, A. (2013). Linked and open data: RDA and bibliographic... | danskin | JILIS.it. *Italian Journal of Library and Information Science*, 4(1), 147-149. doi:10.4403/jlis.it-5463
- De Lama, N. (2012). BIG DATA: Alignment of supply & demand. *PlanetData Roadmapping Workshop with Experts. Proposals for Future Research and Innovations on Big Data*, Retrieved from <http://planet-data.eu/sites/default/files/PlanetData%20Roadmapping.pdf>
- DERI, Fondazione Bruno Kessler & OpenLink software. (2014). Sindice. Retrieved from <http://sindice.com/main/about>
- Deutsche Nationalbibliothek. (2014a). Deutsche nationalbibliothek - home. Retrieved from http://www.dnb.de/DE/Home/home_node.html;jsessionid=ACEB7B407D1A5E2E90D1313DAD66CF20.prod-worker2
- Deutsche Nationalbibliothek. (2014b). Katalog der deutschen nationalbibliothek. Retrieved from <https://portal.dnb.de/opac.htm?method=newSearch¤tView=simple&selectedCategory=any>
- Digital Curation Centre. (2014). What is digital curation? Retrieved from <http://www.dcc.ac.uk/digital-curation/what-digital-curation>
- Dirección General de Modernización Administrativa, Procedimientos e Impulso de la Administración Electrónica. (2013). In Ministerio de Hacienda y Administraciones Públicas (Ed.), *Guía de aplicación de la norma técnica de interoperabilidad de reutilización de recursos de información*. Madrid: Retrieved from http://administracionelectronica.gob.es/pae_Home/dms/pae_Home/documentos/Estrategias/pae_Interoperabilidad_Inicio/Normas tecnicas/Reuse of information resources Interoperability Standar NIF Spain/ENGLISH Interoperability Agreement for%20the%20Reuse%20of%20Information%20Resources.pdf
- Dodds, L. (2013). Open data rights statement vocabulary. Retrieved from <http://schema.theodi.org/odrs/>
- Dublin Core Metadata Initiative. (2014). DCMI home: Dublin core® metadata initiative (DCMI). Retrieved from <http://dublincore.org/>
- Duerst, M., & Suignard, M. (2005). *Internationalized resource identifiers (IRIs) request for comments: 3987* IETF.
- Dunsire, G., Corey, H., Hillman, D., & Phipps, J. (2012). Linked data vocabulary management: Infrastructure, support data, integration and interoperability. *Information Standards Quarterly*, 24(2/3), 4-13. Retrieved from http://www.niso.org/apps/group_public/download.php/9422/isqv24no2-3.pdf

- Dunsire, G., & Willer, M. (2013). *Bibliographic information organization in the semantic web* (<http://proquest.safaribooksonline.com/book/web-applications-and-services/9781843347316?bookview=overview> ed.). Oxford: Chandos Publishing. Retrieved from <http://proquest.safaribooksonline.com/book/web-applications-and-services/9781843347316/copyright/b9781843347316500077.htm>
- Dunsire, G. (2012). Representing the FR family in the semantic web. *Cataloging & Classification Quarterly*, 50(5-7), 724-741. doi:10.1080/01639374.2012.679881
- Escolano-Rodríguez, E. (2013). ISBD adaptation to semantic web of bibliographic data in linked data. *Italian Journal of Library and Information Science*, 4(1), 120-137. doi:10.4403/jlis.it-5484
- Euclid. (2013). EUCLID educational curriculum for the usage the linked data. Retrieved from <http://www.euclid-project.eu/modules/chapter2>
- European Commission. (2013). Guidelines on data management in horizon 2020. *The EU Framework Programme for Research and Innovation HORIZON 2020*, Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- European Commission. (2014a). Commission launches pilot to open up publicly funded research data. Retrieved from http://europa.eu/rapid/press-release_IP-13-1257_en.htm
- European Commission. (2014b). ICT-enabled public sector innovation in H2020. Retrieved from <http://ec.europa.eu/programmes/horizon2020/en/news/ict-enabled-public-sector-innovation-h2020>
- Feigenbaum, L., & Prud'hommeaux, E. (2014). SPARQL by example (W3C SPARQL working group). Retrieved from W3C SPARQL Working Group
- Ferrer-Sapena, A., & Sánchez-Pérez, E. A. (2013). Open data, big data: ¿hacia dónde nos dirigimos? *Anuario ThinkEPI*, 7, 150-156. Retrieved from <http://www.thinkepi.net/open-data-big-data-hacia-donde-nos-dirigimos#sthash.pL3BgzvM.dpuf>
- Fons, T., Penka, J., & Wallis, R. (2012). OCLC's linked data initiative: Using schema.org to make library data relevant on the web. *Information Standards Quarterly*, 24(2/3), 29-35.
- Food and Agriculture Organization of the United Nations. (2014). AGROVOC. Retrieved from <http://aims.fao.org/standards/agrovoc/about>
- Ford, K. (2012). LC's bibliographic framework initiative and the attractiveness of linked data. *Information Standards Quarterly*, 24(2/3), 46-50. doi:10.3789/isqv24n2-3.2012

- Fox, M. S. (2013). City data: Big, open and linked. *Department of Mechanical and Industrial Engineering University of Toronto*, Retrieved from <http://eil.utoronto.ca/smartcities/papers/City-Data-v4.pdf;:>
- Freed, N. (2014). IANA media types. Retrieved from <http://www.iana.org/assignments/media-types/media-types.xhtml>
- Gandon, F., & Schreiber, G. (2014). RDF 1.1 XML syntax W3C recommendation. Retrieved from <http://www.w3.org/TR/rdf-syntax-grammar/>
- García, E. (2014). Sobre el concepto de "big data". *IV Congreso Nacional De Interoperabilidad Y Seguridad*, Retrieved from <http://www.cnis.es/images/informes/Articulo%20Big%20Data%200.0.pdf>
- Gartner. (2012). *The importance of 'Big data': A definition.* (). New York: Gartner Inc. Retrieved from <https://www.gartner.com/doc/2057415/importance-big-data-definition>
- Gearon, P., Passant, A. & Polleres, A. (2013). SPARQL 1.1 update W3C recommendation 21 march 2013. Retrieved from <http://www.w3.org/TR/sparql11-update/>
- Geonames. (2014). GeoNames. Retrieved from <http://www.geonames.org/>
- Giaretta, D. (2011). PARSE.insight | permanent access to the records of science in europe. Retrieved from <http://www.parse-insight.eu/>
- Gil, Y., & Miles, S. (2013). PROV model primer. Retrieved from <http://www.w3.org/TR/prov-primer/>
- Gil-Urdiciain, B. (2004). *Manual de lenguajes documentales*. Gijón: Trea.
- Godby, C. (2013). *The relationship between BIBFRAME and OCLC's linked-data model of bibliographic description: A working paper the relationship between BIBFRAME and OCLC's linked-data model of bibliographic description: A working paper - 2013-05.pdf.* (Working Paper No. OCLC (WorldCat): 850705869). Dublin-Ohio: OCLC Research. Retrieved from <http://oclc.org/content/dam/research/publications/library/2013/2013-05.pdf>
- Guerrini, M., & Possemato, T. (2013). Linked data: A new alphabet for the semantic web. *Italian Journal of Library and Information Science*, 4(1), 67-90. doi:10.4403/jlis.it-6305.
- Harpring, P. (2012). Introduction to controlled vocabularies (getty research institute). Retrieved from http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/constructing.html

- Harris, S., & Seaborne, A. (2013). SPARQL 1.1 query language. Retrieved from <http://www.w3.org/TR/sparql11-query/>
- Hartig, O., & Zhao, J. (2010). Publishing and consuming provenance metadata on the web of linked data. In Springer (Ed.), *Provenance and annotation of data and processes* (pp. 78-90)
- HasLhofer, B., & Isaac, A. (2014). Europeana linked open data. data.europeana.eu. *Semantic Web Journal*, 1, 1-7. Retrieved from http://www.semantic-web-journal.net/system/files/swj297_1.pdf
- Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1), 1-136. doi:10.2200/S00334ED1V01Y201102WBE001
- Herman, I., Adida, B., Sporny, M. & Birbeck, M. (2013). RDFa 1.1 primer - second edition rich structured data markup for web documents W3C working group note. Retrieved from <http://www.w3.org/TR/rdfa-primer/>
- Hernández-Pérez, T., & García-Moreno, M. (2013). Datos abiertos y repositorios de datos: Nuevo reto para los bibliotecarios. *El Profesional De La Información*, 22(3), 259-263. doi:dx.doi.org/10.3145/epi.2013.may.10
- Hitzler, P., & Janowicz, K. (2013). Linked data, big data, and the 4th paradigm. *The Semantic Web Journal*, 4(3), 233-245. Retrieved from <http://www.semantic-web-journal.net/content/linked-data-big-data-and-4th-paradigm>
- Hitzler, P., Krösch, M., Parsia, B. & Rudolph, S. (2012). OWL 2 web ontology language primer (second edition) W3C recommendation. Retrieved from <http://www.w3.org/TR/owl2-primer/>
- Howarth, L. C. (2012). FRBR and linked data: Connecting FRBR and linked data. *Cataloging & Classification Quarterly*, 50(5-7), 763-776. doi:10.1080/01639374.2012.680835
- Hyland, B., Ateamezing, G. & Villazón-Terrazas, B. (2014). Best practices for publishing linked data. Retrieved from <http://www.w3.org/TR/ld-bp/>
- Hyvönen, E. (2012a). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1), 1-159. doi:10.2200/S00452ED1V01Y201210WBE003
- Hyvönen, E. (2012b). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1), 1-159. doi:10.2200/S00452ED1V01Y201210WBE003

- Iacono, A. (2013). Verso un nuovo modello di OPAC. dal recupero dell'informazione alla creazione di conoscenza. *Italian Journal of Library and Information Science*, 4(2), 85-107. doi:10.4403/jlis.it-8903
- Iacono, A. (2014). Dal record al dato. linked data e ricerca dell'informazione nell'OPAC. *Italian Journal of Library and Information Science*, 5(1), 78-102. doi:10.4403/jlis.it-9095
- Iannella, R., Guth, S., Paehler, D. & Kasten, A. (2012). ODRL version 2.0 core model. Retrieved from <http://www.w3.org/community/odrl/two/model/>
- IFLA. (2011). Element set, ISBD elements. Retrieved from <http://iflstandards.info/ns/isbd/elements.rdf>
- IFLA. (2014). International standard bibliographic description. Retrieved from <http://www.ifla.org/publications/international-standard-bibliographic-description>
- Instituto centrale per il catalogo unico delle biblioteche italiane, & European Commission. (2014). LINKED HERITAGE coordination of standards and technologies for the enrichment of europeana. Retrieved from <http://www.linkedheritage.eu/>
- International Federation of Library Associations. (2014a). FRBR review group. Retrieved from <http://www.ifla.org/node/794>
- International Federation of Library Associations. (2014b). ISBD linked data study group. Retrieved from <http://www.ifla.org/node/1795>
- International Organization for Standardization (ISO). (2006). *Topic maps ISO/IEC 13250:2003*. Switzerland: International Organization for Standardization.
- International Organization for Standardization (ISO). (2011). *Information and documentation - thesauri and interoperability with other vocabularies - part 1: Thesauri for information retrieval*. Switzerland: ISO 25964 -1:2011 (E).
- International Organization for Standardization (ISO). (2012). *Information and documentation - thesauri and interoperability with other vocabularies - part 2: Interoperability with other vocabularies*. Switzerland: ISO 25964 -2:2012 (E).
- Isaac, A., Clayphan, R., & HasLhofer, B. (2012). Europeana: Moving to linked open data. *Information Standards Quarterly*, 24(23), 34-40. Retrieved from http://www.niso.org/apps/group_public/download.php/9422/isqv24no2-3.pdf
- Isaac, A., & De Smedt, J. (2013). ISO 25964 SKOS extension. Retrieved from <http://pub.tenforce.com/schemas/iso25964/skos-thes/>

- Isaac, A., Meghini, C., Dekkers, M., Gradmann, S., Clayphan, R., Molendijk, J., . . . Van de Sompel, H. (2013). In Europeana-EU (Ed.), *Europeana data model primer*. La Haya: Retrieved from <http://pro.europeana.eu:9580/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5>
- Isaac, A., & Summers, E. (2005). SKOS simple knowledge organization system primer. Retrieved from <http://www.w3.org/TR/skos-primer/>
- Isaac, A., Waites, W., Young, J. & Zeng, M. (2011). Library linked data incubator group: Datasets, value vocabularies, and metadata element sets. W3C incubator group report 25 october 2011. Retrieved from <http://www.w3.org/2005/Incubator/llid/XGR-llid-vocabdataset-20111025/>
- ISO TC46/SC9/WG8 working group for the ISO 25964, & Isaac, A. (2012). Correspondence between ISO 25964 and SKOS/SKOS-XL models. Retrieved from http://www.niso.org/apps/group_public/download.php?document_id=12351
- Joint Steering Committee for Development of RDA. (2014). RDA registry. Retrieved from <http://www.rdaregistry.info/>
- Jones, S., Marieke, G., & Pickton, M. (2013). *Research data management for librarians* Digital Curation Centre. Retrieved from <http://www.dcc.ac.uk/sites/default/files/documents/events/RDM-for-librarians/RDM-for-librarians-booklet.pdf>
- Keith, A., Cyganiak, R., Hausenblas, H. & Zhao, J. (2011). Describing linked datasets with the Void vocabulary. Retrieved from <http://www.w3.org/TR/void/#dublin-core>
- Keyser, P., & Leuven, K. H. (2012). *Indexing: From thesauri to the semantic web (chandos series for information professionals)* Chandos Publishing. Retrieved from <http://proquest.safaribooksonline.com/book/library-and-information-science/9781843342939>
- Kitchin, R. (2013). Four critiques of open data initiatives. Retrieved from <http://www.nuim.ie/progcity/2013/11/four-critiques-of-open-data-initiatives/#comment-128>
- Korn, N., & Oppenheim, C. (2011). *Licensing open data: A practical guide*. (Guía). London: JISC.
- Kroeger, A. (2013). The road to BIBFRAME: The evolution of the idea of bibliographic transition into a post-MARC future. *Cataloging & Classification Quarterly*, 51(8), 873-890. doi:10.1080/01639374.2013.823584

- Le Boeuf, P. (2013). *Customized OPACs on the semantic web: The OpenCat prototype*. (). París: Bibliothèque nationale de France. Retrieved from <http://files.dnb.de/svensson/UILLD2013/UILLD-submission-3-formatted-final.pdf>
- Leroi, M. V., Holland, J., & Cagnot-Dédale, S. (2011). *Your terminology as part of the semantic web: Recommendations for design and management*. (Recommendations). Roma: Linked Heritage WP3 and ATHENA WP4. Retrieved from <http://www.linkedheritage.org/getFile.php?id=244>
- Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno. LeyU.S.C. I. Disposiciones generales (2013).
- Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, I. Disposiciones generales (2007). Retrieved from Boletín Oficial del Estado Retrieved from <http://www.boe.es/buscar/doc.php?id=BOE-A-2007-19814>;
- Library of Congress. (2012a). MADS/RDF documentation. Retrieved from <http://www.loc.gov/standards/mads/rdf/>
- Library of Congress. (2012b). MADS/RDF namespace document. Retrieved from <http://www.loc.gov/standards/mads/rdf/v1.html>
- Library of Congress. (2014a). Bibliographic framework initiative. Retrieved from <http://www.loc.gov/bibframe/>
- Library of Congress. (2014b). LC linked data service (library of congress). Retrieved from <http://id.loc.gov/>
- Library of Congress. (2014c). PREMIS: Preservation metadata maintenance activity. Retrieved from <http://www.loc.gov/standards/premis/>
- Library of Congress, & Zepheria. (2012). *Bibliographic framework as a web of data: Linked data model and supporting services*. (No. 2014). Whashington: Library of Congress. Retrieved from <http://www.loc.gov/bibframe/pdf/marclid-report-11-21-2012.pdf>
- Maali, F., & Erickson, J. (2014). Data catalog vocabulary (DCAT). Retrieved from <http://www.w3.org/TR/vocab-dcat/>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburg, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. (). New York: McKinsey & Company. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

- Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Almasi, E. (2013). *Open data: Unlocking innovation and performance with liquid information*. (). New York: McKinsey Global Institute. Retrieved from http://www.mckinsey.com/~media/McKinsey/dotcom/Insights/Business%20Technology/Open%20data%20Unlocking%20innovation%20and%20performance%20with%20liquid%20information/MGI_Open_data_FullReport_Oct2013.ashx
- Marcos-Martín, C., & Soriano-Maldonado, L. (2011). Reutilización de la información del sector público y open data en el contexto español y europeo. proyecto aporta. *El Profesional De La Información*, 20(3), 291-297. doi:10.3145/epi.2011.may.07
- Márquez Fernández, J. M., Vázquez Martínez, R., Martínez López, M., & Roldán Cruz, N. (2013). *Estudio de la demanda y uso de gobierno abierto en España*. (). Madrid: Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información.
- Martínez-González, M., & Alvite Díez, M. L. (2014). Propuesta metodológica de evaluación de gestores de tesauros compatibles con la web semántica. *Anales De Documentación*, 17(1), 1-18. doi:10.6018/analesdoc.17.1.186271
- Martínez-Urbe, L., & Macdonald, S. (2008). Un nuevo cometido para los bibliotecarios académicos: Data curation. *El Profesional De La Información*, 17(3), 273-280. doi:10.3145/epi.2008.may.03
- Martini, P. (2013). Bibliographic standards and linked data. towards a collaboration between cultural and commercial. *Italian Journal of Library and Information Science*, 4(1), 305-311. doi:10.4403/jlis.it-8598
- McCallum, S. (2012). *Bibliographic framework initiative approach for MARC data as linked data*. Unpublished manuscript. Retrieved 6/10/2014, Retrieved from http://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CDMQFjAB&url=http%3A%2F%2Fwww.loc.gov%2Fbibframe%2Fpdf%2FBFI-IGeLU-ppt2003.pdf&ei=xuiWU-CKD82Y0AWG3IGADw&usq=AFQjCNE9hQxtbvKu5rEz9zir7Pyq9ejA_w&bvm=bv.68445247,d.d2k&cad=rja
- Méndez, E. (2010). Tendencias en recuperación de información: Principios y retos para una nueva década de datos enlazados. *Anuario ThinkEPI*, 4, 231-239.
- Méndez-Rodríguez, E., & Greenberg, J. (2012). Linked data for open vocabularies and HIVE's global framework. *El Profesional De La Información*, 21(3), 236-244. doi:x.doi.org/10.3145/epi.2012.may.03

- Miles, A., & Bechhofer, S. (2009). SKOS eXtension for labels (SKOS-XL) namespace document - HTML variant, 18 august 2009 recommendation edition. Retrieved from <http://www.w3.org/TR/skos-reference/skos-xl.html>
- Miles, A., & Bechofer, S. (2009). SKOS simple knowledge organization system reference. Retrieved from <http://www.w3.org/TR/skos-reference/#schemes>
- Ministerio de Educación, Cultura y Deporte. (2014). Lista de encabezamientos de materia para las bibliotecas públicas en SKOS. Retrieved from <http://id.sgcb.mcu.es/lem/>
- Norma técnica de interoperabilidad de reutilización de recursos de la información, N.T.U.S.C. (2013). Retrieved from BOE Retrieved from <http://www.boe.es/boe/dias/2013/03/04/pdfs/BOE-A-2013-2380.pdf>
- Ministerio de Industria, Energía y Turismo. (2014a). Agenda digital para españa. Retrieved from http://www.agendadigital.gob.es/agenda-digital/recursos/Recursos/1.%20Versi%C3%B3n%20definitiva/Agenda_Digital_para_Espana.pdf
- Ministerio de Industria, Energía y Turismo. (2014b). Datos.gob.es. Retrieved from <http://datos.gob.es/datos/>
- Mitchell, E. (2013). Metadata developments in libraries and other cultural heritage institutions. *Library Linked Data: Research and Adoption Library Technology Reports*, 49(5), 5-10. doi:10.5860/ltr.49n5
- Mitchell, I., & Wilson, M. (2012). *Linked data connecting and exploiting big data*. (No. 3378). London: Fujitsu Services Limited. Retrieved from <http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-linked-data.pdf>
- Moos, P. (2013). Replacing MARC: Where to start. *Information Standards Quarterly*, 25(4), 14-16. doi:dx.doi.org/10.3789/isqv25no4.2013.03
- Moreiro-Gonzalez, J. A. (2011). *Linguagens documentárias e vocabulários semânticos para a web: Elementos conceituais*. Bahia: EDUFBA.
- Morshed, A., & Rittaban, D. (2012). Machine learning based vocabulary management tool. assessment for the linked open data. *International Journal of Computer Applications*, 60(9), 51-58. doi:10.5120/9724-4197
- Moulaison, H. L., & Million, A. J. (2014). The disruptive qualities of linked data in the library environment: Analysis and recommendations. *Cataloging & Classification Quarterly*, 52(4), 367-387. doi:10.1080/01639374.2014.880981

- National Library of the Netherlands, & European Commission. (2014). Europeana. Retrieved from <http://www.europeana.eu/>
- National Science Digital Library. (2014). Open metadata registry. Retrieved from <http://metadataregistry.org/about.html>
- Nina-Alcocer, N., Blasco-Gil, Y., & Peset, F. (2013). Datasharing: Guía práctica para compartir datos de investigación. *El Profesional De La Información*, 22(noviembre-diciembre), 562-568. doi:dx.doi.org/10.3145/epi.2013.nov.09
- OCLC. (2014a). Data strategy and linked data. Retrieved from <http://www.oclc.org/data.en.html>
- OCLC. (2014b). Dewey decimal classification. Retrieved from <http://dewey.info/>
- OCLC. (2014c). VIAF fichero de autoridades virtual internacional. Retrieved from <http://viaf.org/>
- OECD. (2007). In OCDE Publications (Ed.), *Principles and guidelines for access to research data from public funding. paris: OECD publications*. Paris: Retrieved from <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- Open Government Partnership. (2013). Open government partnership. Retrieved from <http://www.opengovpartnership.org/es>
- Open Knowledge Foundation. (2013). Open government data. Retrieved from <http://opengovernmentdata.org/>
- Open Knowledge Foundation. (2014a). Ckan - the open source data portal software. Retrieved from <http://ckan.org/>
- Open Knowledge Foundation. (2014b). Data packages - data protocols. Retrieved from <http://dataprotocols.org/data-packages/>
- Open Knowledge Foundation. (2014c). Welcome - the datahub. Retrieved from <http://datahub.io/en/>
- Oren, E., Delbru, R., Catasta, M., Cyganiak, R., & Tummarello, G. (2008). Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3, 1-22. doi:10.1.1.136.3952
- Directiva 95/46/CE del parlamento europeo y del consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, (1995).
- Directiva 96/9/CE del parlamento europeo y del consejo, de 11 de marzo de 1996, sobre la protección jurídica de las bases de datos, DirectivaU.S.C. (1996).

- DIRECTIVA 2013/37/UE del parlamento europeo y del consejo de 26 de junio de 2013 por la que se modifica la directiva 2003/98/CE relativa a la reutilización de la información del sector público, DirectivaU.S.C. (2013). Retrieved from Eur-Lex Retrieved from <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:ES:PDF>
- Pastor-Sánchez, J. A. (2013a).
Ampliando skos a partir de la norma de tesauros ISO 25964. *Anuario ThinkEPI*, 7, 189-193.
- Pastor-Sánchez, J. A. (2013b).
Marcado semántico: Tecnologías y aplicación para la representación de sistemas de organización del conocimiento en el contexto linked open data. *Scire*, 19(2), 55-68.
- Peset, F., Ferrer-Sapena, A., & Subirats-Coll, I. (2011). Open data y linked open data: Su impacto en el área de bibliotecas y documentación. *El Profesional De La Información*, 20(2), 165-163. doi:DOI: 10.3145/epi.2011.mar.0
- Picco, P., & Ortiz-Repiso, V. (2012a). RDA, el nuevo código de catalogación: Cambios y desafíos para su publicación. *Revista Española De Documentación Científica*, 35(1), 145-173. doi:10.3989/redc.2012.1.848
- Picco, P., & Ortiz-Repiso, V. (2012b). The contribution of FRBR to the identification of bibliographic relationships: The new RDA-based ways of representing relationships in catalogs. *Cataloging & Classification Quarterly*, 50(5-7), 622-640. doi:10.1080/01639374.2012.680847
- Real decreto 1495/2011, de 24 de octubre, por el que se desarrolla la ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal. Real decretoU.S.C. I. Disposiciones generales (2011).
- RECOLECTA. Grupo de Trabajo de Depósito y Gestión de datos en Acceso Abierto. (2012). La conservación y reutilización de los datos científicos en España. fundación española para la ciencia y la tecnología, FECYT. Retrieved from http://www.fecyt.es/fecyt/detalle.do?elegidaSiguiente=&elegidaNivel3=;SalaPrensa;publicaciones;guiasymanuales&elegidaNivel2=;SalaPrensa;publicaciones&elegidaNivel1=;SalaPrensa&tc=publicaciones&id=Informe_preliminar_datos_cientificos
- Rodriguez, V., Poveda-Villalón, M., Suarez, M. C. & Gomez, A. (2013). Linked data rights. Retrieved from <http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/>
- Rodriguez-Doncel, V., Gomez-Pérez, A., & Mihindukulasooriya, N. (2013). Rights declaration in linked data. *Proceedings of the 4th international workshop on consuming linked data (COLD2013)* (pp. 1-13). Sydney, Australia: CEUR-WS.

- Sánchez-Cuadrado, S., Morato-Lara, J., Moreiro-Gonzalez, J. A., & Marrero-Linares, M. (2007). Definición de una metodología para la construcción de sistemas de organización del conocimiento a partir de un corpus documental en lenguaje natural. *Procesamiento Del Lenguaje Natural*, 39, 213-220. Retrieved from <http://hdl.handle.net/10045/3015>
- Saorín, T., Peset, F., & Ferrer-Sapena, A. (2013). Factores para la adopción de *linked data* e implantación de la web semántica en bibliotecas, archivos y museos. *Information Research*, 18(1) Retrieved from <http://InformationR.net/ir/18-1/paper570.html>
- Schema.org. (2012). Getting started with schema.org. Retrieved from <https://schema.org/>
- Schreibe, G., & Raimond, Y. (2014). RDF 1.1 primer. W3C working group note. Retrieved from <http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/#example1>
- Semantic Web Company GmbH. (2012). How to access members of an RDF list with SPARQL. Retrieved from 2014
- Semantic Web Company GmbH. (2013). PoolParty user guide. Retrieved from <https://grips.semantic-web.at/display/public/POOLDOKU/PPT++User+Guide>
- Semantic Web Company GmbH. (2014). PoolParty semantic suite. Retrieved from <http://www.poolparty.biz/>
- Seventh Framework Programme FP7/2013. (2014). PRELIDA project. Retrieved from <http://www.prelida.eu/>
- Shieh, J. (2013). A transformative opportunity: BIBFRAME at the george washington university, an early experimenter. *Information Standards Quarterly*, 25(4), 17-22. doi:10.3789/isqv25no4.2013
- Shiri, A. (2014). Linked data meets big data: A knowledge organization systems perspective. *Advances in Classification Research Online*, 24(1), 16-20. doi:10.7152/acro.v24i1.14672
- Sigit-Sayogo, D., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, 30, 19-31. doi:dx.doi.org/10.1016/j.giq.2012.06.011
- Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. (2014). JSON-LD 1.0 a JSON-based serialization for linked data W3C recommendation

. Retrieved from <http://www.w3.org/TR/json-ld-syntax/#iris>
- Subirats, I., & Zeng, M. (2013). LOD-BD recommendations 2.0 : How to select appropriate encoding strategies for producing linked open data (LOD)-enabled bibliographic data. Retrieved from <http://aims.fao.org/es/lode/bd>

- Summers, E., Isaac, A., Redding, C., & Krech, D. (2008). LCSH, SKOS and linked data. Paper presented at the *DCMI '08 Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, Berlin., 2014(7/15/2014) 25-33.
doi:10.1016/j.websem.2013.05.001
- Svensson, L. (2013). Are current bibliographic models suitable for integration with the web? *Information Standards Quarterly*, 25(4), 6-13. Retrieved from http://www.niso.org/apps/group_public/download.php/11942/isqv25no4.pdf.pdf
- Swedish University. (2014a). LIBRIS. Retrieved from <http://libris.kb.se/?language=en>
- Swedish University. (2014b). Librisbloggen. Retrieved from <http://librisbloggen.kb.se/2008/12/03/libris-available-as-linked-data/>
- Tascón, M. (2013). Introducción a big data: Pasado, presente y futuro. *Telos. Cuadernos De Comunicación E Innovación*, 95(Junio-Septiembre), 46-50. Retrieved from http://www.fundacion.telefonica.com/es/arte_cultura/publicaciones/detalle/260#
- Tauberer, J. (2007). The 8 principles of open government data. Retrieved from <http://opengovdata.org/>
- The Royal Society. (2012). *Science as an open enterprise*. London: The Royal Society Science Policy Centre. Retrieved from https://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
- Tillet, B. (2013). RDA and the semantic web, linked data environment. *Italian Journal of Library and Information Science*, 4(1), 139-144. doi:<http://dx.doi.org/10.4403/ilis.it-6303>
- Torres-Salinas, D. (2010). Compartir datos (data sharing) en ciencia: Contexto de una oportunidad. *Anuario ThinkEPI*, 4, 258-261. Retrieved from <http://www.thinkepi.net/compartir-datos-data-sharing-en-ciencia-el-contexto-de-una-oportunidad>
- Torres-Salinas, D., Robinson-García, N., & Cabezas Clavijo, A. (2012). Compartir los datos de investigación: Introducción al data sharing. *El Profesional De La Información*, 21(2), 173-184. doi:dx.doi.org/10.3145/epi.2012.mar.08
- Torres-Salinas, D., Martín-Martín, A., & Fuente-Gutiérrez, E. (2014). Analysis of the coverage of the data citation index – thomson reuters: Disciplines, document types and repositories. *Revista Española De Documentación Científica*, 37(1), 1-6.
doi:dx.doi.org/10.3989/redc.2014.1.1114

- Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. *OECD Working Papers on Public Governance*, (22)
doi:<http://dx.doi.org/10.1787/5k46bj4f03s7-en>
- UDC Consortium. (2014). UDC summary. Retrieved from
<http://www.udcc.org/udcsummary/php/index.php>
- Universidad de Gotinga. (2013). In Guibault L., & Wiebe A. (Eds.), *Safe to be open. study on the protection of research data and recommendations for access and usage*. Gotinga: Universidad de Gotinga. Retrieved from
<http://webdoc.sub.gwdg.de/univerlag/2013/legalstudy.pdf>
- University of Mannheim, OPENLIK & Universität Leipzig. (2014). DBpedia. Retrieved from
<http://dbpedia.org/About>
- University of North Texas. (2014). Denton declaration: An open data manifesto. Retrieved from
<http://openaccess.unt.edu/denton-declaration>
- Vanderfeesten, M. (2012). Use case enhanced publications - library linked data wiki. Retrieved from
http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Enhanced_Publications
- Van-Hooland, S., Verborgh, R., & Van-de-Valle, R. (2012). Joining the linked data cloud in a cost-effective manner. *Information Standards Quarterly*, 24(2/3), 22-28. Retrieved from
http://www.niso.org/apps/group_public/download.php/9423/IP_VanHooland-et-al_%20LD-Cloud_isqv24no2-3.pdf
- Vatant, B. (2013). Vocabulary of a friend (VOAF). Retrieved from
<http://lov.okfn.org/vocab/voaf/v2.3/index.html>
- Vesse, R., Hall, W., & Carr, L. (2010). Preserving linked data on the semantic web by the application of link integrity techniques from hypermedia. *Linked data on the web (LDOW2010)* (). Raleigh, NC: Retrieved from <http://eprints.soton.ac.uk/270897/>
- Vila-Suero, D. (2011). Library linked data incubator group: Use cases. Retrieved from
<http://www.w3.org/2005/Incubator/lld/XGR-lld-usecase-20111025/>
- Vila-Suero, D. (2014). Datos.bne.es 2.0: Datos enlazados en la biblioteca nacional de españa. Retrieved from <http://www.slideshare.net/DanielVilaSuero/datosbnees-20>
- Vila-Suero, D., Villazón-Terrazas, B., & Gómez-Pérez, A. (2012). Datos.bne.es. A library linked dataset. *Semantic Web Journal*, 1(6), 1-7.
- Villanova University's Falvey Memorial Library. (2014). Vufind, search, discover, share. Retrieved from <http://vufind-org.github.io/vufind/>

- Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Perez, A. (2011). Methodological guidelines for publishing government linked data

 boris villazón-terrazas, luis. M. vilches-blázquez, oscar corcho, asunción gómez-pérez *Linking government data* (http://rd.springer.com/chapter/10.1007/978-1-4614-1767-5_2 ed., pp. 27-49). New York: Spriger. doi:10.1007/978-1-4614-1767-5_2
- W3C. (2014). W3C semantic web: Search engines. Retrieved from http://www.w3.org/2001/sw/wiki/Category:Search_Engine
- Willer, M., Dunsire, G., & Bosancic, B. (2012). ISBD and the semantic web. *Italian Journal of Library and Information Science*, 1(2), 213-236. doi:10.4403/jlis.it-4536
- Wonderlich, J. (2010). Ten principles for opening up government information. Retrieved from <http://sunlightfoundation.com/policy/documents/ten-open-data-principles/>
- Wood, D., Zaidman, M., Ruth, L., & Hausenblas, H. (2014). *Linked data: Structured data on the web* (ed.). New York: Manning Publications. Retrieved from <http://proquest.safaribooksonline.com/book/web-development/9781617290398>